nature mental health

Article

https://doi.org/10.1038/s44220-024-00229-x

Predicting suicidality with small sets of interpretable reward behavior and survey variables

Received: 4 April 2023

Accepted: 6 March 2024

Published online: 09 May 2024

Check for updates

Shamal Lalvani[1], Sumra Bari[2,8], Nicole L. Vike[2,8], Leandros Stefanopoulos[1,3,8], Byoung-Woo Kim[2,8], Martin Block[4], Nicos Maglaveras[3], Aggelos K. Katsaggelos[1,5,6,9] & Hans C. Breiter[2,7,9]

The prediction of suicidal thought and behavior has met with mixed results. This study of 3,476 de-identified participants (4,019 before data exclusion) quantified the prediction of four suicidal thought and behavior (STB) variables using a short reward/aversion judgment task and a limited set of demographic and mental health surveys. The focus was to produce a simple, quick and objective framework for assessing STB that might be automatable, without the use of big data. A balanced random forest classifier performed better than a Gaussian mixture model and four standard machine learning classifiers for predicting passive suicide ideation, active suicide ideation, suicide planning and planning for safety. Accuracies ranged from 78% to 92% (optimal area under the curve between 0.80 and 0.95) without overfitting, and peak performance was observed for predicting suicide planning. The relative importance of features for prediction showed distinct weighting across judgment variables, contributing between 40% and 64% to prediction per Gini scores. Mediation/moderation analyses showed that depression, anxiety, loneliness and age variables moderated the judgment variables, indicating that the interaction of judgment with mental health and demographic indices is fundamental for the high-accuracy prediction of STB. These findings suggest the feasibility of an efficient and highly scalable system for suicide assessment, without requiring psychiatric records or neural measures. The findings suggest that STB might be understood within a cognitive framework for judgment with quantitative variables whose unique constellation separates passive and active suicidal thought (ideation) from suicide planning and planning for safety.

Suicide rates in the United States increased by over 30% between 2000 and 2020[1], and these rates were exacerbated by the COVID-19 pandemic[2,3]. Efforts predicting the potential for suicidal action are mixed, with some researchers being critical of prediction accuracy[4,5]. Recent research suggests that machine learning (ML) algorithms outperform traditional statistical approaches for the prediction of suicidal thought and behavior (STB)[6,7]. Furthermore, meta-analysis suggests that theories of suicide (for example, biosocial, biological, ideation and hopelessness theories[8]) perform suboptimally when compared to ML algorithms in the prediction of suicidal ideation, suicidal attempt(s) and completed

A full list of affiliations appears at the end of the paper. e-mail: breitehs@ucmail.uc.edu

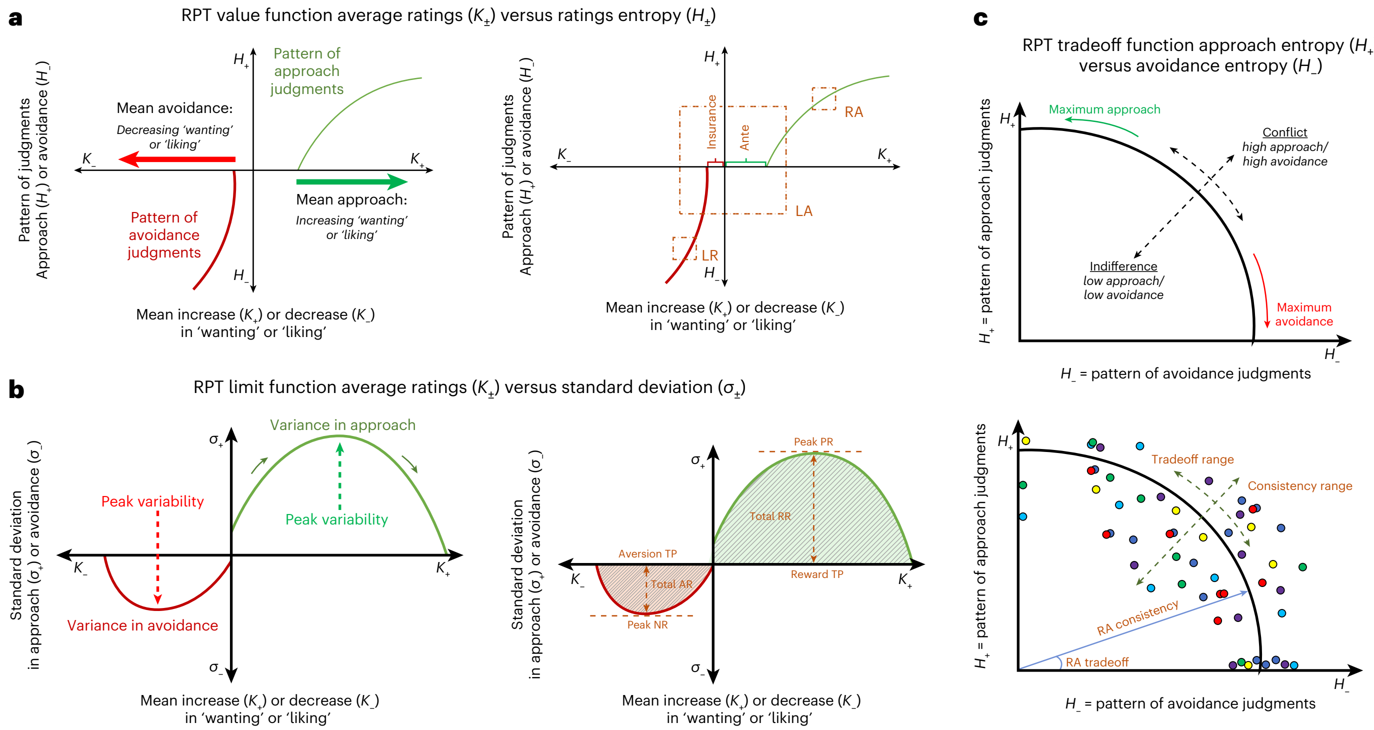

Fig. 1 | Description of judgment variables from Relative Preference Theory. **a**, The value function in relative preference theory (RPT) resembles the value function in Kahneman and Tversky's prospect theory (PT)[43], albeit with very different variables. In both RPT and PT, functions follow a concave power-law function, and the avoidance curve tends to have a greater slope than the approach curve. The entropy H of ratings for each category is a function of the mean rating K for each stimulus category used in the rating task. **b**, The limit function in RPT corresponds to the variance–mean curve produced in Markowitz's portfolio theory[105], which shows the risk level an individual is willing to accept for a fixed reward. The variance of the picture rating of a category (σ) is plotted against the mean rating K for each category of images, producing a parabolic relationship in individuals. **c**, The tradeoff function in RPT characterizes an individual's pattern of approach judgments versus avoidance judgments. The pattern/entropy of approach judgments (H_+) is plotted against the pattern/entropy of avoidance judgments (H_-) for each category of images used in the rating task.

suicide[8]. Grounded by research suggesting that electronically delivered self-report questionnaires correlate significantly with clinical assessment for psychiatric conditions[9], researchers have recently called for the development of a scalable detection platform for the prediction of STB[10].Clinically related measures are predictive of STB, such as post traumatic stress disorder[11] and measures of anger[12]. Additionally, recent research suggests that social and behavioral measures play a key role in the prediction of STB, sometimes surpassing clinical variables in terms of predictive accuracy, such as when used in the context of social media behavior derived from natural language processing (NLP) analysis[10,13] and measures of social integration[14]. This is consistent with other literature suggesting that a clinically valid signal of psychiatric conditions may be available from social media behavior[15]. Although contextual risk factors are not typically studied in ML applications of STB[7], predictive variables of STB are typically contextual[7,16]. For example, substance use and alcohol disorders play a greater risk in suicidal outcomes for veterans and service members than for the general population[16].

Few ML studies have used emergency room (ER)-related questions to predict suicide risk, and no studies have applied small sets (for example, 20–30) of interpretable variables that can be easily acquired on digital devices to predict a set of suicidal thought and behavior variables with high accuracy (Supplementary Fig. 1). STB assessments in the ER generally ask about passive ideation as a framework for opening the topic to query, move to active ideation and planning for harm, and then assess the potential to plan for safety[17]. In the ML literature, passive and active suicidal ideation are often not segregated[18–20], and intent for self-harm and past suicidal attempts tend to be targeted[21–23] rather than plans for suicide. However, truthful responses may be difficult to acquire regarding suicidal ideation and past suicide attempts due to cultural norms or other personal reasons[24,25]. Furthermore, no studies have predicted planning for safety, which is key for framing suicide risk[26], and although it does not reduce suicidal ideation[27], meta-analysis suggests it may reduce relative suicide risk by up to 57% (ref. 27), as well as reduce symptoms of depression, feelings of hopelessness and the incidence of hospitalization[28], although there is a large heterogeneity of suicide planning intervention and study design[28]. No studies have assessed all four STB variables together (that is, passive ideation, active ideation, suicide planning, planning for safety).

Individuals with STB show alterations in reward/aversion judgment or preference[29], such as heightened aversion to risk and loss[30], lower focus on the negative consequences of decisions[31], discounting of delayed rewards[32], and higher bias to escape aversive situations[33]. In economic settings, preferences can be measured from forced choice data (typically through axioms of revealed preference[34]). In the psychological literature, preferences can reflect 'liking' versus 'wanting'[35–39], where the assessment of reward or aversion (that is, judgment) precedes an actual choice. Abnormalities in reward/aversion judgment have been linked to dopamine dysfunction in major depressive disorder, addiction, anxiety, chronic stress[40,41] and STB[41]. Reward/aversion judgment has been mathematically characterized by computational behavior variables reflecting biases[42,43], like loss aversion[44] and risk aversion[45]. Recently, a broader set of 15 variables were found to model unique features of judgment from a picture-rating task that can be implemented on any cellphone or digital device[37,39] (Fig. 1, Supplementary Fig. 2 and Table 1), and these are considered to reflect psychological 'liking'[35,37,39,46,47]. For the present study we hypothesized that this small set of judgment variables (as opposed to big data) would efficiently predict STB.

Table 1 | Description of judgment variables

Preference variable	Abbreviation	Description
(a) Judgment variables derived from the value function or (K, H) curve		
Loss aversion	LA	The degree to which one overweights negative stimuli (or losses) compared to positive stimuli (or gains)
Risk aversion	RA	The degree to which one prefers an uncertain high value outcome to something certain but lower in value
Loss resilience	LR	The extent to which one prefers accepting an uncertain loss to certain loss; it is like RA, but in the domain of losses
Ante	Ante	What one is willing to pay to enter a game of chance (for example, poker)
Insurance	Insurance	The amount of security one is willing to acquire to avoid negative outcomes
(b) Judgment variables derived from the limit function or (K, σ) curve		
Peak positive risk	Peak PR	Per Markowitz's decision utility equation, this is the peak risk around approach choices that must be overcome for approach behavior to occur
Peak negative risk	Peak NR	Per Markowitz's decision utility equation, this is the peak risk around avoidance choices that must be overcome for avoidance behavior to occur
Reward tipping point	Reward TP	Per Markowitz's decision utility equation, this is the reward value beyond which approach choices are made
Aversion tipping point	Aversion TP	Per Markowitz's decision utility equation, this is the intensity of aversion beyond which avoidance choices are made
Total reward risk	Total RR	Total value of reward across the range of risks associated with those positive outcomes; this is the area under the positive variance–mean curve
Total aversion risk	Total AR	The total amount of aversion across the range of risks associated with those negative outcomes; this is the area under the negative variance–mean curve
(c) Judgment variables derived from the tradeoff function or (H_+, H_-) curve		
Reward–aversion tradeoff	RA tradeoff	This represents the balance between approach versus avoidance behavior; it is the mean polar angle between patterns in positive assessments (H_+) and patterns in negative assessments (H_-)
Tradeoff range	Tradeoff range	The variance or bias towards approach versus avoidance behavior; it is one metric of the range in a person's portfolio of preference
Reward–aversion consistency	RA consistency	A continuum between how much an individual has conflict in their reward–aversion preference versus indifference in their preference, where conflict means they both like and dislike something, and indifference means they do not like or dislike something
Consistency range	Consistency range	How much a person swings between conflict and indifference in their preferences; it is a second metric regarding the range in a person's portfolio of preference

Abbreviations and descriptions of relative preference theory (RPT) features. The value function curve derived from (K,H) tuples gives rise to five preference variables: LA, RA, LR, ante and insurance. The limit function curve, derived from (K,σ) tuples, gives rise to six preference variables: peak PR, peak NR, reward TP, aversion TP, total RR and total AR. The tradeoff function, derived from (H_+,H_-) tuples, gives rise to four preference variables: RA tradeoff, tradeoff range, RA consistency and consistency range. The three columns represent the full term for each of the 15 features, their abbreviations as used in this manuscript, and their description. To read about Markowitz's decision utility equation, see ref. 105.

This study tested whether these four STB variables could be predicted with 15 reward/aversion judgment variables (henceforth 'judgment variables', Fig. 1) derived from a short behavior task (Methods). We combined the behavior task with five other survey variables that were hypothesized to contextually frame the judgment variables, but only added minutes to the survey time. This framework was found to make highly accurate prediction using a small, interpretable variable set in lieu of the hundreds to thousands of variables used in traditional big data approaches. Mediation/moderation analyses further revealed that interactions between judgment and survey variables underpinned these high accuracies.

Given the current rates of STB, this approach using limited judgment and survey measures suggests a low-cost approach to STB assessment that could be administered to 85% of the world's population with a personal digital device[48]. Use of variables that do not directly reference STB might also aide identification of at-risk individuals who might be hesitant to disclose self-harm. Ultimately, the power of psychological constructs depends on their capacity to make meaningful predictions, and not just their associations to neural measures.

Results

Adults (ages 18–70 years) across the United States were surveyed in December 2021. High-quality data from 3,476 participants was drawn from (1) Patient Health Questionnaire 8 (PHQ8; absent the question on suicidality[49]), (2) the State Trait Anxiety Inventory–State (STAI)[50], (3) perceived loneliness (self-report), (4) prior attempts at self-harm in the past 1–12 months, (5) five demographic variables known to affect human neuroscience studies (age, ethnicity, education level, sex and handedness)[51–54], (6) 15 judgment variables computationally derived from a simple picture-rating task[37,39] (Supplementary Figs. 3 and 4 and Table 1) and (7) four questions about passive ideation (STB1), active ideation (STB2), suicide planning (STB3) and planning for safety (STB4) on a five-point Likert scale (collectively referred to as STB variables)[17,55] (Methods). The predictive power of variables in (1) to (6) was tested using a balanced random forest (BRF)[56] approach (Supplementary Fig. 5) and Gaussian mixture models (GMMs)[57] (Supplementary Fig. 6) to discriminate between the low and high thresholds of the four STB variables. To provide a baseline against these analyses, we also performed the following four standard ML analyses: random forest (RF), logistic regression (LR), neural network (NN) and support vector machine (SVM)[58]. Given potential personal reluctance or cultural norms[59,60] against reporting past self-harm, variables from (1) to (3) and (5) and (6) were initially tested, followed by a minimal predictor set of (4) to (6). The full set of (1) to (6) was further tested. The relative importance of features used in prediction was evaluated using mutual information (MI) scoring (where a higher MI for a feature and predictor suggests

Table 2 | BRF prediction of STB (*t*=2)

	Accuracy	Sensitivity	Specificity	AUC
(a) Feature set for predicting STB1 (passive ideation)				
RPT	0.619	0.533	0.633	0.614
RPT+PHQ8	0.756	0.813	0.747	0.849
RPT+PHQ8+STAI	0.767	0.841	0.755	0.861
RPT+PHQ8+STAI+loneliness	0.785	0.846	0.775	0.877
RPT+PHQ8+STAI+loneliness+demo.	0.789	0.855	0.778	0.883
RPT+prior attempts (between 1 and 12 months ago)	0.834	0.594	0.873	0.768
RPT+loneliness+prior attempts (between 1 and 12 months ago)	0.804	0.756	0.812	0.851
RPT+loneliness+prior attempts (between 1 and 12 months ago)+demo.	0.816	0.773	0.823	0.869
Entire feature set	0.813	0.851	0.807	0.907
(b) Feature set for predicting STB2 (active ideation)				
RPT	0.627	0.579	0.632	0.647
RPT+PHQ8	0.753	0.844	0.743	0.849
RPT+PHQ8+STAI	0.754	0.852	0.744	0.855
RPT+PHQ8+STAI+loneliness	0.780	0.864	0.772	0.883
RPT+PHQ8+STAI+loneliness+demo.	0.793	0.867	0.785	0.895
RPT+prior attempts (between 1 and 12 months ago)	0.880	0.712	0.897	0.849
RPT+loneliness+prior attempts (between 1 and 12 months ago)	0.864	0.786	0.872	0.900
RPT+loneliness+prior attempts (between 1 and 12 months ago)+demo.	0.874	0.815	0.880	0.908
Entire feature set	0.844	0.872	0.841	0.935
(c) Feature set for predicting STB3 (planning for suicide)				
RPT	0.668	0.610	0.673	0.698
RPT+PHQ8	0.766	0.806	0.762	0.854
RPT+PHQ8+STAI	0.763	0.843	0.756	0.861
RPT+PHQ8+STAI+loneliness	0.797	0.861	0.791	0.891
RPT+PHQ8+STAI+loneliness+demo.	0.811	0.874	0.806	0.905
RPT+prior attempts (between 1 and 12 months ago)	0.914	0.802	0.924	0.898
RPT+loneliness+prior attempts (between 1 and 12 months ago)	0.905	0.828	0.912	0.934
RPT+loneliness+prior attempts (between 1 and 12 months ago)+demo.	0.902	0.838	0.908	0.942
Entire feature set	0.888	0.887	0.889	0.953
(d) Feature set for predicting STB4 (planning for safety)				
RPT	0.717	0.715	0.718	0.604
RPT+PHQ8	0.710	0.732	0.707	0.764
RPT+PHQ8+STAI	0.730	0.743	0.728	0.766
RPT+PHQ8+STAI+loneliness	0.799	0.576	0.834	0.786
RPT+PHQ8+STAI+loneliness+demo.	0.738	0.753	0.735	0.796
RPT+prior attempts (between 1 and 12 months ago)	0.777	0.641	0.799	0.737
RPT+loneliness+prior attempts (between 1 and 12 months ago)	0.619	0.533	0.633	0.791
RPT+loneliness+prior attempts (between 1 and 12 months ago)+demo.	0.785	0.667	0.804	0.814
Entire feature set	0.777	0.741	0.783	0.831

Accuracy, sensitivity, specificity and area under the curve (AUC) for the BRF prediction of STB1–STB4. A threshold of 2 indicated that a scores of 1 or 2 for each STB variable were compared to scores of 3–5 for each STB variable on the five-point Likert-like scale. A score of 1 or 2 indicated none to mild levels of the STB variable, whereas 3–5 indicated higher levels of the STB variable. In terms of group numbers, 2,991 participants had a score of 1 or 2, 485 had scores of 3–5 for STB1, 3,155 participants had a score of 1 or 2, 321 had scores of 3–5 for STB2, 3,209 participants had a score of 1 or 2, 267 had scores of 3–5 for STB3, and 3,004 participants had a score of 1 or 2, 472 had scores of 3–5 for STB4. Feature sets first start with judgment variables (RPT) and successively include the PHQ8 score, the STAI score, loneliness and demographics, respectively. Analyses were also conducted using a feature set with only judgment variables and prior attempts at hurting oneself, followed by successively adding loneliness and demographic (demo.) variables (age, gender at birth, ethnicity, education level and handedness). Finally, all features were included for the prediction of each STB variable. Values were rounded off at three decimal places.

predictive power)[61] and Gini score plots[62]. Mechanistic relationships between the top predictors were assessed using statistical mediation and moderation, where the four STB variables were dependent variables.

Prediction of STB variables

Given the higher sensitivity and specificity of BRF analyses, BRF outcomes are presented in the main text and the GMM results, along with the four standard ML results, are provided in Supplementary

Tables 10–19. In the following sections, ML results for judgment variables, PHQ8 score, STAI score and loneliness are described in the first paragraph, and the results for judgment variables, prior attempts and loneliness are described in the second paragraph. Results with inclusion of all predictors (judgment variables, PHQ8 score, STAI score, loneliness and prior attempts) are described in the third paragraph.

Passive suicidal ideation (STB1). BRF prediction of STB1 (rated on a Likert scale of 1–5, where 1 = no suicidal ideation and 2–5 = increasing degrees of suicidal ideation; that is, threshold = 1) using judgment variables yielded higher accuracies with PHQ8, STAI and loneliness variables included (59.0–78.8%) (Supplementary Table 1a). Sensitivities and specificities improved from 51.4% to 83.3% and from 61.0% to 77.7%, respectively. Adding demographics improved these metrics by less than 2%. Fusion of the PHQ8 score with judgment variables led to a consistent boost of ~18% for accuracy and 32% for sensitivity. Results with inclusion of judgment, PHQ8, STAI and loneliness features were similar (61.9–78.5%; Table 2a) when the threshold for passive suicidal ideation was set to 2.

When judgment variables were fused with reports of prior suicide attempts and loneliness, predictive accuracy of STB1 at threshold = 1 was 78.4% and 78.1%, respectively. Sensitivity improved from 52.0% to 74.2% when loneliness was fused with judgment variables and prior attempts, whereas specificity showed a decrease from 85.4% to 79.1%. Prediction of STB1 at threshold = 2 when judgment variables were fused with reports of prior suicide attempts and loneliness was 83.4% and 80.4%, respectively, with similar sensitivities and specificities to threshold = 1.

Analysis with all predictors achieved maximum AUC scores of 0.905 for STB1 threshold = 1, and 0.907 when threshold = 2, achieving sensitivities of 84.8% and 85.1%, respectively.

Active suicidal ideation (STB2). BRF prediction of STB2 (threshold = 1) using judgment variables yielded higher accuracies as PHQ8, STAI and loneliness variables were successively included (63.8–78.7%; Supplementary Table 1b). Sensitivities and specificities improved from 56.0% to 86.1% and 65.1% to 77.5%, respectively. Further adding demographics improved these metrics by less than 1%. Fusion of the PHQ8 score with judgment variables boosted measures by 12% for accuracy and 28% for sensitivity. For threshold = 2, judgment variables yielded higher accuracies as PHQ8, STAI and loneliness features were successively included (62.7–78.0%; Table 2b), with similar outcomes for sensitivity and specificity.

Prediction accuracy of STB2 (threshold = 1) when judgment variables were fused with reports of prior suicide attempts and loneliness was 86.4% and 84.7%, respectively. Sensitivity and specificity percentages were in the high 60–70s and 80–90s, respectively. Prediction of active suicidal ideation (threshold = 2) when judgment variables were fused with reports of prior suicide attempts and loneliness was 88.0% and 86.4%, respectively. Sensitivity and specificity percentages were in the 70s and high 80s, respectively.

Analysis with all predictor variables achieved maximum AUC scores of 0.935 at threshold = 1 and 0.931 at threshold = 2, achieving sensitivities in each case of 87.2% and 86.6%, respectively.

Suicide planning (STB3). BRF prediction of STB3 (threshold = 1) using judgment variables yielded higher accuracies as PHQ8, STAI and loneliness variables were successively fused with them (64.4–79.4%; Supplementary Table 1c). Sensitivities and specificities improved from 57.5% to 84.9% and 65.3% to 78.7%, respectively. Further adding demographics improved these metrics by less than 2%. Fusion of the PHQ8 score with judgment variables consistently boosted prediction by ~10% for accuracy and 20% for sensitivity. For threshold = 2, the judgment variables yielded higher accuracies as PHQ8, STAI and loneliness features were successively fused with them (66.8–79.7%; Table 2c).

Prediction accuracy of STB3 (threshold = 1) when judgment variables were fused with reports of prior suicide attempts and loneliness variables was 92.2% and 90.8%, respectively. Sensitivities and specificities were in the high 70s and low 90s. Prediction of STB3 (threshold = 2) when judgment variables were fused with reports of prior suicide attempts and loneliness was 91.4% and 90.5%, respectively. Sensitivities and specificities were in the low 80s and low 90s, respectively.

Analysis with all predictor variables achieved maximum AUC scores of 0.953 with threshold = 1 and 0.948 for threshold = 2, achieving sensitivities of 86.4% and 88.7%, respectively.

Planning for safety (STB4). BRF prediction of STB4 (threshold = 1) using judgment variables yielded higher accuracies as PHQ8, STAI and loneliness features were successively fused with them (59.4–73.8%; Supplementary Table 1d). Further adding demographics improved these metrics by less than 2%. Sensitivities and specificities improved from 55.0% to 74.8% and 60.3% to 73.6%, respectively. Fusion of the PHQ8 score with judgment variables consistently boosted the prediction by ~14% for accuracy and 15% for sensitivity. For threshold = 2, the judgment variables also yielded higher accuracies as PHQ8, STAI and loneliness features were fused with them (71.7–79.9%; Table 2d), with similar outcomes for other metrics.

Prediction accuracy of STB4 (threshold = 1) using judgment variables fused with prior suicide attempts and loneliness variables was 81.0% and 79.2%, respectively (Table 2d). Sensitivity and specificity percentages were in the high 50s and mid 80s, respectively. Prediction of planning for safety (threshold = 2) when judgment features were fused with prior suicide attempts and loneliness was 77.7% and 61.9%, respectively (Table 2d). Sensitivities and specificities were in the 50–60% range and 60–80% range, respectively.

Analysis with all predictor variables achieved maximum AUC scores of 0.837 at threshold = 1 and 0.831 at threshold = 2, while maintaining sensitivities of 73.1% and 74.1%, respectively.

Variable contributions to STB prediction

Distinct sets of judgment variables contributed to prediction of the four STB measures, as measured through normalized MI scoring (Fig. 2 and Supplementary Table 6a). LA had zero MI for each STB measure. For passive suicidal ideation, no judgment variable predominated by MI value, and three had zero-value MIs. This profile was different for active suicidal ideation, where variables for aversion TP and tradeoff range had the highest MI. For suicide planning, the MI with the tradeoff range was far more than other judgment variables. For planning for safety, variables for aversion TP and reward TP were predominant. Despite the distinct patterns of MI for the 15 judgment variables among the four STB measures, regressions between these variables had consistent valences (except reward TP and total AR; Supplementary Table 6b). On the basis of the valence between each judgment variable and STB measure, passive suicidal ideation and suicide planning shared the same patterns, and both differed from active suicidal ideation and planning for safety.

Gini score plots revealed that some survey variables were consistently highest in importance, but the full set of judgment variables were consistently grouped together (Fig. 3 and Supplementary Figs. 7–29). In all analyses, the grouped judgment variables produced summed Gini scores of 0.404 to 0.638—the highest summed Gini scores in 14 of the 24 analyses (Table 3). The 15 judgment variables were consistently more important than education, race, gender and handedness variables.

The rank ordering of Gini scores for the survey variables was distinct for each STB measure, as it was for the 15 judgment variables. Despite this, STAI measures tended to have one of the top two Gini scores, and age was consistently one of the bottom of the five survey variables.

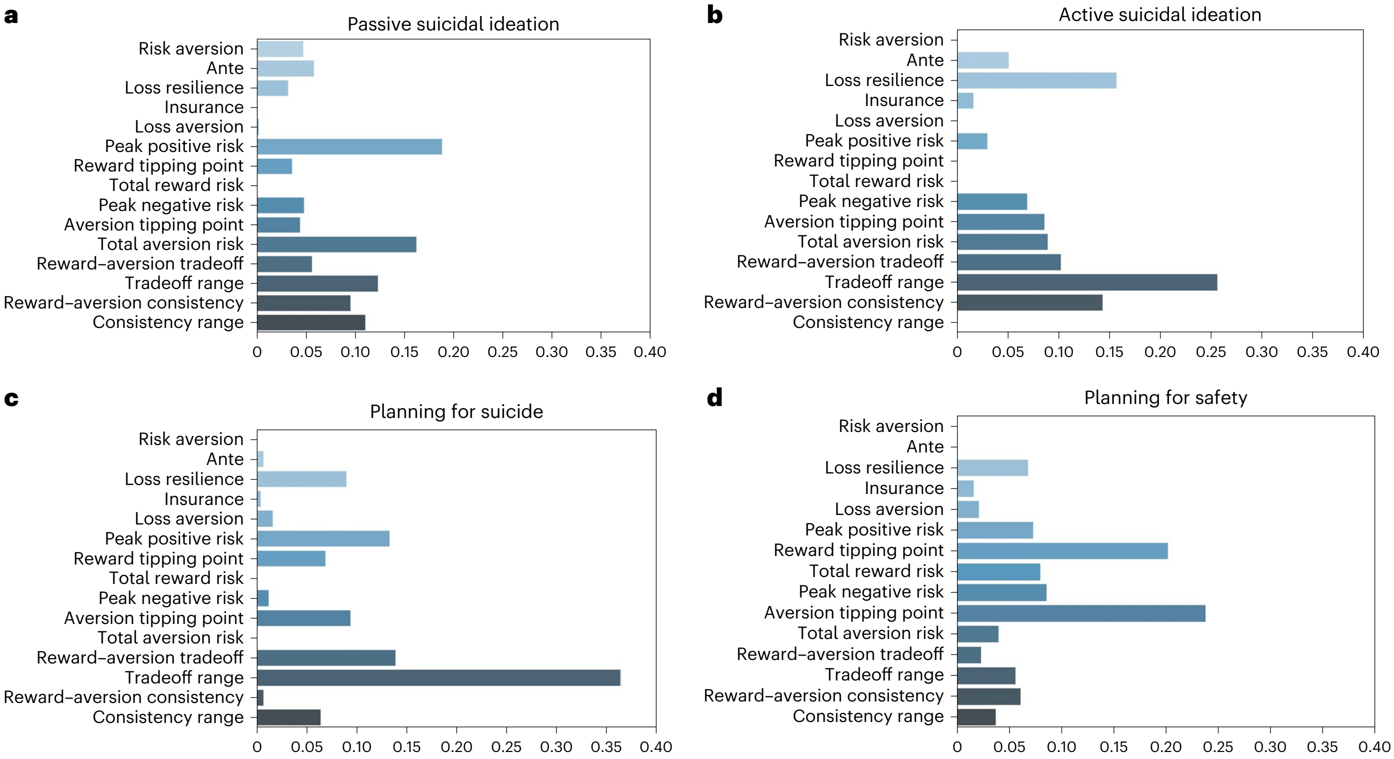

Fig. 2 | Mutual information scores between RPT features and STB variables. **a**–**d**, MI scoring of RPT features with respect to passive suicidal ideation (**a**), active suicidal ideation (**b**), planning for suicide (**c**) and having a plan for safety (**d**). Exact MI values are listed in Supplementary Table 6a. The MI between two variables informally expresses the amount of information gained about one variable by observation of another. In this context, this relates to the amount of information gained about STB variables by knowledge of the judgment variables. The length of the bars in the figures represents the MI (*x* axis) of the RPT variables (*y* axis). Longer bars indicate a higher MI between RPT variables and STB. This alludes to a larger predictive value. A MI score of zero implies that the two variables are independent, and therefore that prediction of one variable based on another is unlikely.

a

Variable	Mean decrease in Gini coefficient
Prior attempts	0.138
Age	0.09
Loneliness	0.087
Consistency range	0.047
Peak positive risk	0.045
Ante	0.045
Total reward risk	0.045
Reward tipping point	0.044
Loss resilience	0.044
Tradeoff range	0.044
Reward–aversion consistency	0.043
Risk aversion	0.043
Total aversion risk	0.041
Aversion tipping point	0.041
Peak negative risk	0.04
Loss aversion	0.04
Reward–aversion tradeoff	0.039
Insurance	0.037
Education	0.023
Race	0.01
Gender	0.009
Handedness	0.005

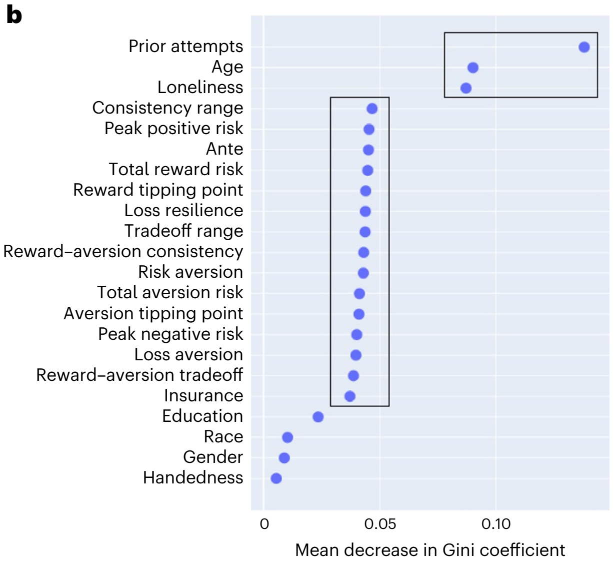

Fig. 3 | Gini importance scores for judgment and contextual variables for planning for safety. **a**, Gini importance values of the entire set of features used to predict planning for safety (STB4) at $t = 2$ using a BRF. The numerical values displayed are the mean decrease in Gini coefficient that occurs by removing each variable. The larger the mean decrease in Gini coefficient, the greater importance the feature has in the classifier. **b**, Visual display of the Gini importance values. Note that two boxes are shown. The highest box highlights the top variables in terms of Gini importance values, and the lower box represents the Gini importance values of the judgment variables.

Mediation/moderation analysis

Mediation and moderation analyses followed published approaches[54,63] and were driven by the Gini score analyses[62]. The majority of judgment variables were involved in mediation/moderation relationships ($\alpha = 0.05$), excluding LA, total RR and reward TP for mediation, and excluding total RR and total AR for moderation (Supplementary Tables 2–4). Survey variables statistically mediated the relationship between 11 judgment variables and passive suicidal ideation, whereas they statistically

Table 3 | Gini contribution of judgment variables, top variables and bottom variables

Variable	Threshold	Gini contribution of behavior	Gini contribution of top variables	Gini contribution of bottom variables
(a) Feature set: prior attempts, loneliness, behavior variables and demographics				
STB1	1	0.572	0.384	0.043
	2	0.556	0.403	0.041
STB2	1	0.496	0.470	0.037
	2	0.488	0.479	0.033
STB3	1	0.428	0.536	0.037
	2	0.435	0.531	0.035
STB4	1	0.631	0.321	0.048
	2	0.638	0.315	0.047
(b) Feature set: PHQ8, STAI, behavior variables and demographics				
STB1	1	0.451	0.515	0.035
	2	0.467	0.501	0.034
STB2	1	0.464	0.504	0.033
	2	0.459	0.505	0.036
STB3	1	0.472	0.488	0.041
	2	0.478	0.486	0.038
STB4	1	0.594	0.361	0.044
	2	0.605	0.349	0.045
(c) Feature set: entire feature set				
STB1	1	0.552	0.415	0.032
	2	0.558	0.412	0.030
STB2	1	0.587	0.386	0.027
	2	0.593	0.380	0.026
STB3	1	0.622	0.351	0.032
	2	0.603	0.366	0.030
STB4	1	0.423	0.534	0.040
	2	0.404	0.553	0.041

Summed Gini importance for the full set of judgment variables (15 total), the top non-judgment variables from the Gini importance plots, and the bottom variables (those at the bottom of the Gini importance plots) for STB1–STB4. These are presented for STB variable thresholds of 1 and 2 with the following feature sets: (a) prior attempts, loneliness, behavior variables and demographics, (b) PHQ8, STAI, behavior variables and demographics and (c) the entire feature set of all variables.

moderated the relationship between eight judgment variables and passive suicidal ideation. For the other three STB measures, there was minimal mediation involving perceived loneliness, PHQ8 and STAI survey variables. Instead, there were salient moderation effects for these three survey measures with 12 of the 15 judgment variables (Supplementary Table 4).

Of the five survey variables, prior suicide attempts demonstrated mediation with four judgment variables to predict STB2–4 and moderated three judgment variables to predict STB1–2. Age showed mediation with six judgment variables, and moderation with one judgment variable (Supplementary Table 4).

Discussion

This study sought a short, objective and automatable framework for predicting four STB measures using 15 variables for biases in reward/aversion judgment and a very limited set of demographic and mental health survey indices. Given that reward/aversion judgment is known to be affected by demographic and mental health indices, we fused demographics with the anxiety, depression and judgment variables for ML. This produced five primary results. (1) All four STB metrics were predicted with small sets of predictors within a range of 78–92% accuracy and 0.796–0.953 AUC; this compares favorably with the literature[4,19–21,23,64–70]. (2) Judgment variables and limited survey indices were most effective at predicting planning for suicide, producing accuracies in the upper range of what other studies have reported for suicide risk and suicide attempts (for example, AUC = 0.857 and 0.99, respectively, in the literature and 0.953 here) without using complex big data approaches (for example, 100+ variables or inclusion of neuroimaging) or retrospective data[19,20,64–70]. (3) Prediction of active suicidal ideation and planning for suicide was improved by the addition of one self-report measure of prior attempts of self-harm, similar to the addition of depression and other mental health indices. (4) Mediation/moderation analyses showed that depression, anxiety, loneliness and age variables had significant moderation effects on judgment variables, indicating that the interaction of mental health and contextual indices with judgment variables statistically predicted STB. (5) BRF prediction far outperformed GMM prediction and standard ML prediction (that is, RF, LR, NN and SVM prediction), particularly for the sensitivity index.

Collection of the limited set of variables used for prediction was more feasible and less time-consuming when compared to previous studies using larger datasets with hundreds to thousands of variables for prediction[19,20,64–70], and this task can be easily implemented on any digital device[37–39]. When the other mental health indices and demographics are included, data acquisition takes ~5 min to complete but has comparable prediction results to big data approaches[19,20,64–70]. Inclusion of the prior self-harm variable greatly improved the prediction of active suicidal ideation and planning for suicide, but did not improve prediction of passive suicidal ideation or planning for safety. This suggests that prior history might impact intention for harm, but not its inverse—the planning for self-preservation. Of the other mental health indices, the neurovegetative symptoms of depression in the PHQ8 most improved the predictive accuracy of the four STB variables, although by themselves each mental health index and age had higher Gini scores compared to the judgment variables. Loneliness is commonly considered a risk factor for STB[18], and it was found to moderate the largest number of judgment variables in the prediction of STB variables. The addition of demographic variables was also not consistently beneficial for the prediction of STB, and competition between variables, leading to poorer prediction, cannot be ruled out[71].

The sensitivity metric is important for evaluating the prediction efficacy of STB variables, particularly if intervention might be considered[4,19–21,23,64–70]. According to a review of studies since 2017 that reported a sensitivity metric for predicting suicidal ideation (Supplementary Table 5)[4,19–21,23,64–70], sensitivities ranged from 41% to 87%, with AUC scores ranging from 0.61 to 0.94, in line with our findings[4,19,21,64,65]. Higher AUC scores and sensitivities were typically provided for peer-reviewed work with large feature sets consisting of neuroimaging data, although none of these publications segregated passive and active ideation, as done commonly in clinical interviews[19,64]. Furthermore, a number of studies relied on anonymized electronic health records, and did not explicitly report the number of features used, which can vary per participant[4,21,23,65,67]. A similar range of sensitivities has been reported for the prediction of suicide attempt(s) (30.8–100%) with AUC scores ranging from 0.59 to 0.99 in cohorts of between 75 and 2,959,689 participants[4,19–21,23,65,66]. The prediction of suicidal attempt(s) appears to be more common in the literature than other STB outcomes, but has the caveat of being predominantly retrospective reporting[4,19–21,23,65,66]. In studies in the past five to six years providing sensitivity results, prediction of suicide risk provided similar sensitivities to suicidal ideation and suicidal attempt(s) (between 59% and 85.3%), and AUC scores of up to 0.857 (refs. 68,69). Finally, it must be noted that prediction of completed suicide reports showed much lower sensitivities (between 28% and 69%) and lower AUC scores (between 0.66 and 0.8)[21,70]. Relative to this literature, prior studies have not (1) segregated passive from active ideation in prediction of suicidal ideation[4,19,21,64,65], (2) explicitly predicted planning for harm or (3) explicitly predicted planning for safety, all while achieving results in the upper range of what is reported for prediction of suicidal ideation, suicidal attempts and completed suicides (Supplementary Table 5)[4,19–21,23,64–70].

The type of ML used for imbalanced data has become a topic of substantial research[72–75]. In this study, BRF prediction outperformed GMM prediction and four standard approaches to ML, consistent with the literature[76]. With BRF prediction, each STB variable further had a unique profile of judgment variables that contributed to their prediction, in that the MI metric of variable contribution was unique (Fig. 2 and Supplementary Table 6). The only exception was loss aversion (LA), which had no mutual information with any STB variable, and was not involved in any mediation/moderation relationship. The values for LA observed in this study were quite low, consistent with other work using picture ratings where there is no consequence for making a rating, unlike an operant keypress that changes view time[37,39]. Each of the four STB variables was best classified by a unique weighting of the 15 judgment variables, arguing that distinct aspects of reward/aversion judgment are important for each of the four STB metrics. This observation raises the hypothesis that unique constellations of judgment variables may underlie other forms of mental health conditions and behavior. Should further work show this for other mental health conditions (for example, depression, anxiety, substance use disorder), such findings would strongly support calls from the ML community to develop a standard model of mind[77], albeit a model centered around processes for judgment and agency that focus on reward and aversion assessments by an organism.

Some limitations should be considered. First, the cohort was collected from the United States. As psychopathologies may differ across cultures[24,25], cultural influences may result in different judgment variable groupings affecting prediction. Second, all variables were self-reported and not from clinical records, although it is not clear how a prospective study of STB with thousands of participants performing an experimental cognitive task could be run if there was a chance that breaking the blind would save lives. Third, the cohort was sampled during the COVID-19 pandemic, in which greater incidents of loneliness and suicidality were reported[2,3], arguing for further work in the absence of a pandemic.

Conclusions

The current work found that 15 judgment variables and limited mental health and demographic information predicted four STB measures with sensitivities and specificities around 80% using a BRF approach that produced the highest sensitivities of the approaches used. There appear to be few studies that integrate quantitative judgment features, such as from a short behavioral task, to predict distinct STB measures. This work supports publications suggesting that social and behavioral measures play a key role in the prediction of STB, sometimes surpassing clinical variables in predictive accuracies[10,13]. Contextual risk factors are also not typically studied in ML applications of STB[7], yet predictive variables of suicidality can be contextual[7,16] and the reported mediation/moderation results strongly support these reports. The current results contrast tendencies in the literature to either (1) use large feature sets for prediction (for example, hundreds to thousands of variables)[19,20,64,66–70] or (2) collect expensive clinical or biological measures for prediction[20–23,64,65,67–70]. The data needed for prediction in this study can readily be acquired by smart phones and other digital devices, which are currently available for 92% of the US population[78] and 85% of the world population[48]. The analysis does not require a supercomputer and thus can be scaled to populations for which big data and expensive clinical or biological measures are not available, meeting frameworks proposed by others for the development of a scalable detection platform for prediction of suicidality[10]. By combining multiple variables around STB, including assessments of planning for safety, this framework suggests a digital approach for early assessment and triage, which is particularly needed now[10].

Going forward, future work might assess a broader set of features that can be extracted from the preference curves (that is the value, tradeoff and limit functions) besides the 15 used herein. Additionally, the analysis could be expanded to more deeply assess age and other demographic variables, such as between groups of adolescents and elderly participants. These age groups have different contextual risk factors[79–82], making relevant contextual variables (retirement versus work status, insurance coverage, other medical illnesses, illness in peer group, familial social network and social media usage) potentially relevant variables. The study might also be followed up post pandemic to see if other variables become important as predictors or mediators/moderators. Given the current results, and the fact that every step of data collection, analysis and prediction can be automated, research groups with the requisite expertise might move forward with testing of such a system for populations at high risk (for example, higher education and the military)[83,84].

Methods

Cohort recruitment

A third-party vendor, Gold Research Inc., recruited a population sample of adults across the United States (ages 18–70 years), with a final sample of 4,019 participants (see refs. 85,86 for the recruitment framework and procedures). To ensure adequate samples of participants with mental health conditions, Gold Research oversampled 15% of the sample for mental health conditions. Participant demographics were matched to the US Census Bureau at the time of sampling in December 2021. A total of 4,019 adults participated (mean age ± s.d. = 51.4 ± 14.9 years) (full demographics are provided in Supplementary Table 7), and, after applying data exclusion criteria, the data of 3,476 participants were retained. Informed consent was obtained for all participants, which included their primary participation in the study as well as the secondary usage of their anonymized, de-identified (that is, all identifying information removed by Gold Research Inc. before retrieval by the research group) data in secondary analyses. Informed consent was approved by the Institutional Review Board of Northwestern University (STU00213665 (ref. 55)) and the University of Cincinnati (2023-0164), in accordance with the Declaration of Helsinki.

Reward/aversion judgment task

Participants completed a picture-rating task on their personal computers or cellphones. Each participant viewed a randomized sequence of 48 images, displayed one at a time. Images were from the International Affective Picture Set[87], with eight images from each of the six picture categories sports, disasters, cute animals, aggressive animals, nature and adults in bathing suits. Participants were asked to rate each image on an integer scale from −3 (a strong disliking) to +3 (a strong liking), with zero being neutral (Supplementary Fig. 2). There was no time limit for making a picture rating, although participants were asked to rate the images as quickly as possible and to use their first impression. The next image was displayed once a rating was selected. The instructions shown to participants were as follows:

> ‘The next part of this survey involves looking at pictures and then responding how much you like or dislike the image. Please rate each image on a scale from −3 (Dislike Very Much) to +3 (Like Very Much). Zero (0) is neutral… meaning you have no feelings either way. The images are a set of photographs that have been used by scientists around the world for over 20 years.
>
> It is important you rate each picture based on your initial emotional response. There are no right or wrong answers… just respond with your feelings and rate the pictures very quickly.
>
> Please click ‘Next’ to begin.’

See refs. 37,39 for further details.

Mental health indices and demographics

Demographics were acquired for five variables that have established relationships with neuroimaging[51–54]. Two published surveys were used in this study: (1) the Patient Health Questionnaire-9 (ref. 49), with the question about suicide removed and henceforth referred to as PHQ8, and (2) the State Trait Anxiety Inventory (STAI), where only state questions were used[50]. We further queried (1) perceived loneliness (self-report of a five-point Likert-like scale) and (2) the number of prior attempts at harming oneself in the past year. These variables (henceforth 'survey variables'), along with the judgment variables (described below), were inputs for supervised ML prediction. We sought to predict four STB measures adopted from the Massachusetts General Hospital Subjective Question screener (MGH SQ) used in the Phenotype Genotype Project in Addiction and Mood Disorders[55,85,88]: (i) passive suicidal ideation, (ii) active suicidal ideation (STB1), (iii) planning for suicide and (iv) planning for safety. For the variables in (1), (2) and (i)–(iv), we used a five-point Likert scale: 1 being 'Never', 2 being 'Rarely', 3 being 'Sometimes', 4 being 'Often' and 5 being 'Always'. Survey ratings for (i)–(iv) were answered by participants as relating to the past month. The MGH SQ has been used in multiple studies[38,89–94] and the four STB questions had been adapted to the MGH SQ from a clinical textbook on emergency psychiatry[17].

Measurement of passive suicidality ('Wishing to go to sleep and not wake up') corresponded directly to the criteria of passive suicidality in the Columbia Suicide Severity Rating Scale (CSSRS)[95] and the Columbia Lighthouse Project for the Navy (CLPN)[96], which measured passive suicidality as either wishing to be dead or wishing to go to sleep and not wake up. Active suicidality ('Wanting to hurt yourself or take your own life') also corresponded to measurements in the CSSRS and CLPN; however, it did not explicitly measure intent (for example, 'Have you had any thoughts about how you might do this'). Planning for suicide ('Having a plan to take your own life') similarly corresponded to survey questions in the CSSRS (for example, 'Have you thought about doing something to make yourself not alive anymore?') and CLPN (for example, 'Have you done anything, started to do anything, or prepared to do anything to end your own life?'). We note that planning for safety ('Having a safety plan for not hurting yourself when these feelings arise') is not specifically measured in the CSSRS or CLPN. However, it is a fundamental component of assessing suicide risk[17,26].

Specific demographics collected were (1) age group, (2) gender at birth (that is, sex), (3) race/ethnicity, (4) highest education level completed and (5) handedness. Demographic categories and frequencies are listed in Supplementary Table 7.

General data exclusion

The following quality assurance procedures were implemented as employed in other publications[37,39,55,85]. Participants meeting at least one of the following six criteria were omitted from the cohort: (1) participants with ten or more clinician-diagnosed illnesses, (2) participants that selected the same response for at least one section of the survey, (3) participants that rated all images in the behavioral task the same or with a variance of 1 (meaning only two of seven Likert points were used), (4) participants whose relative preference analysis yielded extreme outliers >3 Interquartile Ranges (IQRS) or incomplete measurements, (5) participants that had mismatching responses to years of education and education level in the survey, (6) participants that completed the questionnaire in less than 800 s (refs. 55,85). Data exclusion reduced the sample from 4,019 participants to 3,476 participants for analysis.

Reward/aversion judgment analysis

Computational behavior analysis used code published in refs. 38,39. Ratings (Supplementary Fig. 2) were analyzed for each participant as schematized in Supplementary Fig. 3, to produce relative preference theory (RPT) graphs (Fig. 1), which share a striking similarity to prospect theory and portfolio theory graphs[38,88,97], yet use distinct variables. Procedures were performed in MATLAB as detailed elsewhere[38,88,97] (Supplementary Methods). Relative preference variables were extracted using MATLAB 7.1 with the following toolboxes: Curve Fitting Toolbox 1.1.4, Image Processing Toolbox 5.1, MATLAB Builder for Excel 1.2.5, Statistics Toolbox 5.1 and Symbolic Math Toolbox 3.1.3.

As described in Supplementary Fig. 3, picture ratings produced an average magnitude (K), variance (σ) and pattern or information (that is, Shannon entropy (H)) related to participants' preference behavior. The variable K reflects the average (mean) of the positive ratings (K_+) or negative ratings (K_-) a participant made within each picture category. Similarly, the variance in positive ratings (σ_+) or negative ratings (σ_-), along with the Shannon entropy (that is, information[98]) of positive ratings (H_+) or negative ratings (H_-) were computed for stimuli within each category. The Shannon entropy characterizes the degree of uncertainty across a set of responses[98] and is a core variable in information theory. Given it quantifies the pattern of judgments made to a set of stimuli, it could be considered a memory variable. These variables capture judgments about the magnitude (intensity of rating) and valence of judgment (positive versus negative or approach versus avoidance) to describe relative preferences (Fig. 1)[37–39,97].

For the computation of H, data were screened for cases where $K = 0$ for a given category (that is, cases where the participant made all neutral ratings to neither approach nor to avoid any stimulus in the category). Computation of H for a given picture category requires that $K > 0$, because H computation results in an undefinable $\log_{10}(0/0)$ when $K = 0$. In such cases, H was set to 0 for categories in which the participant rated '0' for all the stimuli.

To fit the models to participants' ratings, the data were further screened for inclusion/exclusion criteria as follows:

1. Valid entropy (H) calculations (as above)
2. Further exclusion of extreme outliers: loss aversion values > 200, resulting in $N = 42$ exclusions; positive quadratic area >100, resulting in $N = 5$ exclusions
3. Sufficient data points to fit the model with a computable R^2 (for example, at least three points for a nonlinear fit)
4. Coherence of model fits between individual and group data. This last criterion required that the curve direction for individual participant fits be consistent with the curve direction of the group-level statistical fits (and boundary envelopes).

Criteria (3) and (4) are necessary operational definitions for quality assurance given the potential for convergence failures with curve fitting. Overall, 3,476 of 4,019 participants met all quality assurance criteria for picture-rating data and survey data.

According to published procedures, six types of model fitting were performed for the rating data: group and individual models for the (K, H) data, (K, σ) data and (H_+, H_-) data. For the group data, we generated group-level data fits along with boundary envelopes (power-law fits and logarithmic fits for group (K, H) data), and quadratic fits for group (K, σ) data to guide the focus of statistical testing based on the power-law fits (K, H) and quadratic fits (K, σ) for individual data. Individual data then followed these fits based on logarithmic and simple power-law fits for individual (K, H) value functions, quadratic fits for individual (K, σ) limit functions, and radial fits for individual (H_+, H_-) tradeoff distributions[38,97]. For this study sample, participants' (K, H) value functions were fit by concave-logarithmic or power-law functions (Supplementary Table 8 and Supplementary Fig. 4) with all R^2 values >0.80 and ranging from 0.84 to 0.96. For the limit functions, concave quadratic fits across participants' (K, σ) data had goodness of fit assessed using the same metrics as for the (K, H) data (Supplementary Table 8), and the R^2 values varied from 0.85 to 0.94. Finally, radial functions were fit to test for tradeoff plots in the distribution of H_- and H_+ values across categories within each participant, as shown in Supplementary Fig. 4. Value (Supplementary Fig. 4a), limit (Supplementary

Fig. 4b) and tradeoff (Supplementary Fig. 4c) functions were plotted for 500 randomly sampled participants out of 3,476 whose data met all quality assurance criteria. The location and dispersion estimates of the R^2, adjusted R^2 (accounting for degrees of freedom) and associated F-statistics for each participant's model fit are provided in Supplementary Table 8.

From these graphs, judgment variables (Supplementary Table 9) were derived for ML. At least 15 features can be mathematically derived from this framework that are psychologically interpretable and have been validated[37,39] as being discrete, recurrent and scalable by engineering criteria[38,97]. These 15 features are loss aversion, risk aversion, loss resilience, ante, insurance, peak positive risk, peak negative risk, reward tipping point, aversion tipping point, total reward risk, total aversion risk, reward–aversion tradeoff, tradeoff range, reward–aversion consistency and consistency range, as schematized in Fig. 1 and described in Table 1. Loss aversion, risk aversion, loss resilience, ante and insurance are derived from the logarithmic or power-law fit of mean key presses (K) versus the entropy of key presses (H); this is referred to as the value function (Fig. 1a). Peak positive risk, peak negative risk, reward tipping point, aversion tipping point, total reward risk and total aversion risk are derived from the quadratic fit of K versus the standard deviation of key presses (σ); this is referred to as the limit function (Fig. 1b). Reward aversion tradeoff, tradeoff range, reward–aversion consistency, and consistency range are derived from the radial fit of the pattern of avoidance judgments (H_-) versus the pattern of approach judgments (H_+); this is referred to as the tradeoff function (Fig. 1c). Each feature describes a quantitative component of a participant's approach/avoidance or judgment/behavior (see Supplementary Methods for complete descriptions). Collectively, the 15 RPT features are referred to as 'judgment variables' herein.

Feature importance

Feature importance was assessed in two frameworks, one via an MI[61] assessment and the other via a Gini importance score[62] assessment. For this study, MI was computed between each of the 15 judgment variables and the four prediction outcomes of interest (passive suicidal ideation (STB1), active suicidal ideation (STB2), planning for suicide (STB3) and planning for safety (STB4)). The MI was used as a proxy for the dependence between the two random variables, with a larger mutual information between judgment variables and STB metrics suggesting higher importance of the feature in predicting STB. MI scores for the 15 judgment variables are expected to add to 1 for each of the STB measures tested. Note that the STB1–4 nomenclature is used only for figures and tables. Analyses were conducted using the package sklearn[99] to calculate the mutual information scores in Python.

Gini importance scores reflected the feature importance (that is, rank) as determined using the function model.feature_importances_ in Python, and the results were reported and plotted using matplotlib.

Machine learning

The sensitivity of a binary classifier measures its ability to accurately label a symptom-positive class[75]. Sensitivity plays an important role in STB prediction due to the relevance of correctly identifying STB-positive individuals for potential intervention. However, imbalanced datasets typically result in poor sensitivities when the positive class is substantially less in size compared to the negative class[75]. To enable prediction with high sensitivity, two classifiers with the reported ability to handle data imbalance[56,75,100] were compared to a set of standard classifiers. The BRF[56] and GMM[57] classifiers were thus chosen to predict the four STB measures, and compared to four standard classifiers (that is, random forest (RF), logistic regression (LR), neural network (NN) and support vector machine (SVM))[58]. The ML analyses used the parameters detailed in the following (and referenced in ref. 39). Across the analyses, the code was implemented in Python 3.9 using the packages imblearn 0.0, sklearn 1.2.2, pandas 2.0.2 and pandas 2.0.2. Feature importance (Gini scores and MI scores) were obtained with sklearn 1.2.2. Figures related to variable importance were plotted with seaborn 0.12.2.

The classifiers used the following features: (1) PHQ8 (absent the question on suicidality), (2) STAI, (3) perceived loneliness (self-report), (4) prior attempts at self-harm in the past 1–12 months, (5) five demographic variables (age, ethnicity, education level, sex and handedness) and (6) 15 judgment variables computationally extracted from a simple picture-rating task. The predictive power of variables in (1)–(6) was tested using BRF, GMM and the four standard classifiers to discriminate between the low and high measures of the four STB variables. Each STB measure was partitioned as binary data in two different ways for its 1–5 Likert ratings: 1 versus 2–5 (threshold = 1), and 1,2 versus 3–5 (threshold = 2). In each case, the binary data were analyzed with BRF, GMM and the four standard classifiers. Given potential cultural norms against reporting past self-harm[59,60], variables from (1)–(3), (5) and (6) were initially tested, followed by a minimal predictor set of (4)–(6). As a third framework, the full set of (1)–(6) was also tested. In the analysis of variables from (1)–(3), (5) and (6), judgment variables from (6) were first tested, then the other variables were added incrementally. The same was done with the other two analysis frameworks (that is, using just a minimal predictor set for (4)–(6), and using all variables).

BRF analysis. BRF (Supplementary Fig. 5) was implemented in Python with the package imblearn[100] with tenfold cross-validation. RF classifiers contain an ensemble of decision trees from which majority voting is performed to output a class label, and are typically trained by optimizing a Gini or information score[56,99]. In the BRF approach, an ensemble of 200 trees was constructed, where each bootstrap sample was randomly under-sampled to create a balanced dataset of both classes (that is, 50% of STB-positive and 50% of STB-negative data). Balancing was used for training only, and not for testing within cross-validation. No hyperparameter tuning was performed. Subsequent analysis was performed internally using the sklearn RF package through imblearn. Soft labels were used for majority voting, so that the majority vote was weighted on the probability of the sample belonging to the STB-positive class[56,99,100]. The BRF was trained using the Gini criterion, and no maximum tree depth was used, so nodes expanded until leaves were pure or contained at most one sample. The mean accuracy, sensitivity, specificity and AUC were reported.

GMM and standard classifiers. Analyses for GMM, RF, LR, NN and SVM classification were conducted with sklearn[99] in Python (Supplementary Methods and Supplementary Fig. 6).

Mediation/moderation analysis

Unlike standard assessment of linear associations[101,102], mediation assesses the causal pathway between variables and moderation assesses the interaction between such variables to predict a third variable[54,63,103]. Given the number of associations tested prior to mediation/moderation, and the potential for skewed distributions and outliers in human samples, we integrated Cook's distance outlier analysis with mediation[54,63,103] to protect against false positives and increase the analytic power. Mediation/moderation analyses were conducted in R 4.2.0 with the libraries readxl 1.4.0, MASS 7.3-56 and stats 4.2.0. Mediation/moderation analyses used the code sequences detailed in refs. 54,63.

Mediation analysis. Mediation models suggest that, instead of a direct causal relationship between the independent variable (X) and the dependent variable (Y), there is an intermediary variable (M) so that X influences M, which in turn influences Y. Mediation analyses were conducted for all combinations of STB variables as dependent variables, RPT features as independent variables, and prior attempts, age, loneliness, PHQ8 score and STAI score as mediators. The STB variables involved no thresholding. Beta coefficients and their standard error(s) terms from the following linear regression equations, followed the four-step process of ref. 104, and were used from a regression to calculate Sobel P values and mediation effect percentages (T_{eff}):

Step 1: $Y = \gamma_1 + c(X) + \epsilon_1$

Step 2: $M = \gamma_2 + a(X) + \epsilon_2$

Step 3: $Y = \gamma_3 + c'(X) + b(M) + \epsilon_3$

Step 4: Sobel's test was then used to test if c' was significantly lower than c using the following equation:

$$\text{Sobel } z - \text{score} = \frac{c - c'}{\sqrt{b^2 s_{\mathrm{b}}^2 + a^2 s_a^2}} = \frac{ab}{\sqrt{b^2 s_{\mathrm{b}}^2 + a^2 s_{\mathrm{a}}^2}}$$

Using a standard two-tail z-score table, the Sobel P value was determined from the Sobel z-score, and the mediation effect percentage (T_{eff}) was calculated using

$$T_{\mathrm{eff}} = 100 \times \left[1 - \frac{c'}{c}\right]$$

For mediation to be considered significant, we required that all three regressions between X predicting Y (that is, P_{c}), X predicting M (that is, P_{a}), and M predicting Y (that is, $P_{\mathrm{b'}}$) in Supplementary Table 3 show nominal significance with $P < 0.05$. Significant mediation further required the following: $P_{\mathrm{Sobel}} < 0.05$ and $T_{\mathrm{eff}} > 50\%$, following previous publications[54,63,103]. Secondary mediation analysis was run by switching variables assigned to X and M to see if the mediation effects were directed. For secondary analysis if $P_{\mathrm{Sobel}} > 0.05$ and $T_{\mathrm{eff}} < 50\%$, this added to the evidence of M lying in the causal pathway between X and Y.

Moderation analysis. Moderation models suggest that a moderator variable (Mo) controls the magnitude of the relationship between the independent variable (X) and the dependent variable (Y). Moderation analyses were conducted for all combinations of STB variables as dependent variables, RPT features as independent variables, and prior attempts, age, loneliness, PHQ8 score and STAI score as moderators. No thresholds were considered for the STB variables. The original data for STB variables, involving five categories of severity, were used. The moderation analyses involved fitting a logistic regression to the data, described by

$$\log(\mathrm{odds}) = \beta_0 + \beta_1 X + \beta_2 \mathrm{Mo} + \beta_3\,(X \times \mathrm{Mo}) + \epsilon$$

The standard approach was used, where the moderation was deemed significant if the P value of the interaction term ($P_{\ 3}$) and the P value of the overall model (P_{overall}) were both less than or equal to 0.05 through likelihood ratio tests. The likelihood ratio test for the full model was implemented with the following null and alternative hypotheses:

$$H_0 : \beta_1 = \beta_2 = \beta_3 = 0$$

$$H_{\mathrm{A}} : \beta_{\mathrm{i}} \neq 0; \text{for at least one } \beta_{\mathrm{i}}; \text{where } i = 1, 2, 3$$

The likelihood ratio test for the β_3 coefficient was implemented with the following null and alternative hypothesis, respectively:

$$H_0 : \beta_3 = 0$$

$$H_{\mathrm{A}} : \beta_3 \neq 0$$

where the restricted model is

$$\log(\mathrm{odds}) = \beta_0 + \beta_1 X + \beta_2 \mathrm{Mo} + \epsilon$$

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this Article.

Data availability

Data were de-identified before being provided to the investigators. Data are available in Microsoft Excel format and include relative preference variables, demographic metrics and survey variables inclusive of suicidal thought and behavior (STB) variables. The data may be accessed at https://osf.io/6e2d4/. For questions concerning data availability, please contact the corresponding author.

Code availability

Computational behavior analysis used code published in refs. 37,38. ML analyses used parameters as detailed in the Methods and referenced in ref. 39. Mediation/moderation analyses used code sequences as detailed in refs. 54,63.

References

1. *Suicide Data and Statistics* (Centers for Disease Control and Prevention & National Center for Health Statistics); https://www.cdc.gov/suicide/suicide-data-statistics.html
2. Farooq, S., Tunmore, J., Wajid Ali, M. & Ayub, M. Suicide, self-harm and suicidal ideation during COVID-19: a systematic review. *Psychiatry Res.* **306**, 114228 (2021).
3. Hill, R. M. et al. Suicide ideation and attempts in a pediatric emergency department before and during COVID-19. *Pediatrics* **147**, e2020029280 (2021).
4. McHugh, C. M., Corderoy, A., Ryan, C. J., Hickie, I. B. & Large, M. M. Association between suicidal ideation and suicide: meta-analyses of odds ratios, sensitivity, specificity and positive predictive value. *BJPsych Open* **5**, e18 (2019).
5. McHugh, C. M. & Large, M. M. Can machine-learning methods really help predict suicide? *Curr. Opin. Psychiatry* **33**, 369–374 (2020).
6. Linthicum, K. P., Schafer, K. M. & Ribeiro, J. D. Machine learning in suicide science: applications and ethics. *Behav. Sci. Law* **37**, 214–222 (2019).
7. Schafer, K. M. *The Status of Suicidality Prediction Research: A Meta-Analysis*. MSc thesis, Florida State Univ. (2019).
8. Schafer, K. M., Kennedy, G., Gallyer, A. & Resnik, P. A direct comparison of theory-driven and machine learning prediction of suicide: a meta-analysis. *PLoS ONE* **16**, e0249833 (2021).
9. Arrow, K. et al. Evaluating the use of online self-report questionnaires as clinically valid mental health monitoring tools in the clinical whitespace. *Psychiatr. Q.* **94**, 221–231 (2023).
10. Resnik, P., Foreman, A., Kuchuk, M., Musacchio Schafer, K. & Pinkham, B. Naturally occurring language as a source of evidence in suicide prevention. *Suicide Life Threat Behav.* **51**, 88–96 (2021).
11. Schafer, K. M., Clancy, K. J. & Joiner, T. An investigation into the bidirectional relationship between post-traumatic stress disorder and suicidal ideation: a nine year study. *J. Anxiety Disord.* **85**, 102510 (2022).
12. Schafer, K. M. et al. The relationship between anger and suicidal ideation: investigations in two samples. *J. Clin. Psychol.* **78**, 1866–1877 (2022).
13. Shing, H.-C., Resnik, P. & Oard, D. W. A prioritization model for suicidality risk assessment. In *Proc. 58th Annual Meeting of the Association for Computational Linguistics* (ed. Jurafsky, D. et al.) 8124–8137 (Association for Computational Linguistics, 2020).
14. Schafer, K. M., Wilson, E. & Joiner, T. Traumatic brain injury and suicidality among military veterans: the mediating role of social integration. *J. Affect. Disord.* **338**, 414–421 (2023).
15. Kelly, D. L. et al. Can language use in social media help in the treatment of severe mental illness? *Curr. Res. Psychiatry* **1**, 1–4 (2021).
16. Schafer, K. M. et al. Suicidal ideation, suicide attempts and suicide death among Veterans and service members: a comprehensive meta-analysis of risk factors. *Mil. Psychol.* **34**, 129–146 (2022).

17. Hyman, S. & Tesar, G. *Manual of Psychiatric Emergencies* (Little Brown, 1994).
18. Bennardi, M. et al. Longitudinal relationships between positive affect, loneliness and suicide ideation: age-specific factors in a general population. *Suicide Life Threat Behav.* **49**, 90–103 (2019).
19. Chen, V. C.-H. et al. Convolutional neural network-based deep learning model for predicting differential suicidality in depressive patients using brain generalized q-sampling imaging. *J. Clin. Psychiatry* **82**, 19m13225 (2021).
20. Nordin, N., Zainol, Z., Mohd Noor, M. H. & Lai Fong, C. A comparative study of machine learning techniques for suicide attempts predictive model. *Health Informatics J.* **27**, 146045822198939 (2021).
21. Bernert, R. A. et al. Artificial intelligence and suicide prevention: a systematic review of machine learning investigations. *Int. J. Environ. Res. Public Health* **17**, 5929 (2020).
22. Zalar, B., Kores Plesnicar, B., Zalar, I. & Mertik, M. Suicide and suicide attempt descriptors by multimethod approach. *Psychiatr. Danub.* **30**, 317–322 (2018).
23. Rozek, D. C. et al. Using machine learning to predict suicide attempts in military personnel. *Psychiatry Res.* **294**, 113515 (2020).
24. Kirmayer, L. J. The politics of diversity: pluralism, multiculturalism and mental health. *Transcult. Psychiatry* **56**, 1119–1138 (2019).
25. Bredström, A. Culture and context in mental health diagnosing: scrutinizing the DSM-5 revision. *J. Med. Humanit.* **40**, 347–363 (2019).
26. Moscardini, E. H. et al. Suicide safety planning: clinician training, comfort and safety plan utilization. *Int. J. Environ. Res. Public Health* **17**, 6444 (2020).
27. Nuij, C. et al. Safety planning-type interventions for suicide prevention: meta-analysis. *Br. J. Psychiatry* **219**, 419–426 (2021).
28. Ferguson, M., Rhodes, K., Loughhead, M., McIntyre, H. & Procter, N. The effectiveness of the safety planning intervention for adults experiencing suicide-related distress: a systematic review. *Arch. Suicide Res.* **26**, 1022–1045 (2022).
29. Takahashi, T. Neuroeconomics of suicide. *NeuroEndocrinol. Lett.* **32**, 400–404 (2011).
30. Baek, K. et al. Heightened aversion to risk and loss in depressed patients with a suicide attempt history. *Sci. Rep.* **7**, 11228 (2017).
31. Hadlaczky, G. et al. Decision-making in suicidal behavior: the protective role of loss aversion. *Front. Psychiatry* **9**, 116 (2018).
32. Dombrovski, A. Y. et al. The temptation of suicide: striatal gray matter, discounting of delayed rewards and suicide attempts in late-life depression. *Psychol. Med.* **42**, 1203–1215 (2012).
33. Millner, A. J. et al. Suicidal thoughts and behaviors are associated with an increased decision-making bias for active responses to escape aversive states. *J. Abnorm. Psychol.* **128**, 106–118 (2019).
34. Mas-Collel, A., Whinston, M. & Greej, J. *Microeconomic Theory* (Oxford Univ. Press, 1995).
35. Dai, X., Brendl, C. M. & Ariely, D. Wanting, liking and preference construction. *Emotion* **10**, 324–334 (2010).
36. Lee, S. et al. The commonality of loss aversion across procedures and stimuli. *PLoS ONE* **10**, e0135216 (2015).
37. Azcona, E. A. et al. Discrete, recurrent and scalable patterns in human judgement underlie affective picture ratings. Preprint at https://arxiv.org/abs/2203.06448 (2022).
38. Kim, B. W. et al. Recurrent, robust and scalable patterns underlie human approach and avoidance. *PLoS ONE* **5**, e10613 (2010).
39. Vike, N. et al. Predicting COVID-19 vaccination uptake using a small and interpretable set of judgment and demographic variables: Cross-Sectional Cognitive Science Study. *JMIR Public Health Surveill*. https://doi.org/10.2196/47979 (2024).
40. Dillon, D. G. et al. Peril and pleasure: an RDoC-inspired examination of threat responses and reward processing in anxiety and depression. *Depress. Anxiety* **31**, 233–249 (2014).
41. Bogdan, R. & Pizzagalli, D. A. Acute stress reduces reward responsiveness: implications for depression. *Biol. Psychiatry* **60**, 1147–1154 (2006).
42. Tversky, A. & Kaheman, D. Judgment under uncertainty: heuristics and biases. *Science* **185**, 1124–1131 (1974).
43. Kahneman, D. & Tversky, A. Prospect theory: an analysis of decision under risk. *Econometrica* **47**, 263–292 (1970).
44. Chen, X., Voets, S., Jenkinson, N. & Galea, J. M. Dopamine-dependent loss aversion during effort-based decision-making. *J. Neurosci.* **40**, 661–670 (2020).
45. Wang, S., Krajbich, I., Adolphs, R. & Tsuchiya, N. The role of risk aversion in non-conscious decision making. *Front. Psychol.* **3**, 50 (2012).
46. Aharon, I. et al. Beautiful faces have variable reward value: fMRI and behavioral evidence. *Neuron* **32**, 537–551 (2001).
47. Montague, P. R. Free will. *Curr. Biol.* **18**, R584–R585 (2008).
48. *Mobile Fact Sheet* (Pew Research Center, 2021).
49. Spitzer, R. L. Validation and utility of a self-report version of PRIME-MD: the PHQ Primary Care Study. *JAMA* **282**, 1737–1744 (1999).
50. Spielberger, C. D., Gorsuch, R. L., Lushene, R., Vagg, P. R. & Jacobs, G. A. *Manual for the State-Trait Anxiety Inventory* (Consulting Psychologists Press, 1983).
51. HATTA, T. Handedness and the brain: a review of brain-imaging techniques. *Magn. Reson. Med. Sci.* **6**, 99–112 (2007).
52. Dotson, V. M. & Duarte, A. The importance of diversity in cognitive neuroscience. *Ann. N. Y. Acad. Sci.* **1464**, 181–191 (2020).
53. Kaczkurkin, A. N., Raznahan, A. & Satterthwaite, T. D. Sex differences in the developing brain: insights from multimodal neuroimaging. *Neuropsychopharmacology* **44**, 71–85 (2019).
54. Bari, S. et al. Integrating multi-omics with neuroimaging and behavior: a preliminary model of dysfunction in football athletes. *Neuroimage Rep.* **1**, 100032 (2021).
55. Woodward, S. F. et al. Anxiety, post-COVID-19 syndrome-related depression and suicidal thoughts and behaviors in COVID-19 survivors: cross-sectional study. *JMIR Form. Res.* **6**, e36656 (2022).
56. More, A. S. & Rana, D. P. Review of random forest classification techniques to resolve data imbalance. In *Proc. 2017 1st International Conference on Intelligent Systems and Information Management* (*ICISIM*) 72–78 (IEEE, 2017); https://doi.org/10.1109/ICISIM.2017.8122151
57. Hand, D. J., McLachlan, G. J. & Basford, K. E. Mixture models: inference and applications to clustering. *Appl. Stat.* **38**, 384–385 (1989).
58. Watt, J., Borhani, R. & Katsaggelos, A. *Machine Learning Refined* (Cambridge Univ. Press, 2020); https://doi.org/10.1017/9781108690935
59. Aggarwal, S., Borschmann, R. & Patton, G. C. Tackling stigma in self-harm and suicide in the young. *Lancet Public Health* **6**, e6–e7 (2021).
60. Oexle, N. et al. Mental illness stigma, secrecy and suicidal ideation. *Epidemiol. Psychiatr. Sci.* **26**, 53–60 (2017).
61. Duncan, T. E. On the calculation of mutual information. *SIAM J. Appl. Math.* **19**, 215–220 (1970).
62. Breiman, L., Friedman, J. H., Olshen, R. A. & Stone, C. J. *Classification And Regression Trees* (Routledge, 2017); https://doi.org/10.1201/9781315139470
63. Vike, N. L. et al. A preliminary model of football-related neural stress that integrates metabolomics with transcriptomics and virtual reality. *iScience* **25**, 103483 (2022).
64. Weng, J.-C. et al. An autoencoder and machine learning model to predict suicidal ideation with brain structural imaging. *J. Clin. Med.* **9**, 658 (2020).

65. Wilimitis, D. et al. Integration of face-to-face screening with real-time machine learning to predict risk of suicide among adults. *JAMA Netw. Open* **5**, e2212095 (2022).
66. Hettige, N. C. et al. Classification of suicide attempters in schizophrenia using sociocultural and clinical features: a machine learning approach. *Gen. Hosp. Psychiatry* **47**, 20–28 (2017).
67. Macalli, M. et al. A machine learning approach for predicting suicidal thoughts and behaviours among college students. *Sci. Rep.* **11**, 11363 (2021).
68. García de la Garza, Á., Blanco, C., Olfson, M. & Wall, M. M. Identification of suicide attempt risk factors in a national US survey using machine learning. *JAMA Psychiatry* **78**, 398–406 (2021).
69. Sawhney, R., Joshi, H., Gandhi, S., Jin, D. & Shah, R. R. Robust suicide risk assessment on social media via deep adversarial learning. *J. Am. Med. Inform. Assoc.* **28**, 1497–1506 (2021).
70. Kessler, R. C. et al. Predicting suicides after outpatient mental health visits in the Army Study to Assess Risk and Resilience in Servicemembers (Army STARRS). *Mol. Psychiatry* **22**, 544–551 (2017).
71. Bolón-Canedo, V., Sánchez-Maroño, N. & Alonso-Betanzos, A. *Feature Selection for High-Dimensional Data* (Springer, 2015); https://doi.org/10.1007/978-3-319-21858-8
72. Kaur, H., Pannu, H. S. & Malhi, A. K. A systematic review on imbalanced data challenges in machine learning. *ACM Comput. Surv.* **52**, 1–36 (2020).
73. Tran, T., Le, U. & Shi, Y. An effective up-sampling approach for breast cancer prediction with imbalanced data: a machine learning model-based comparative analysis. *PLoS ONE* **17**, e0269135 (2022).
74. Korkmaz, S. Deep learning-based imbalanced data classification for drug discovery. *J. Chem. Inf. Model.* **60**, 4180–4190 (2020).
75. He, H. & Ma, Y. *Imbalanced Learning: Foundations, Algorithms and Applications* (IEEE Press, 2013).
76. Fernandez-Delgado, M., Cernadas, E. & Barro, S. Do we need hundreds of classifiers to solve real world classification problems? *J. Mach. Learn. Res.* **15**, 3133–3181 (2014).
77. Laird, J. E., Lebiere, C. & Rosenbloom, P. S. A standard model of the mind: toward a common computational framework across artificial intelligence, cognitive science, neuroscience and robotics. *AI Mag.* **38**, 13–26 (2017).
78. Malone, L. *Computer and Internet Use in the United States: 2018* (United States Census Bureau, 2021).
79. Wasserman, D., Carli, V., Iosue, M., Javed, A. & Herrman, H. Suicide prevention in childhood and adolescence: a narrative review of current knowledge on risk and protective factors and effectiveness of interventions. *Asia Pac. Psychiatry* **13**, e12452 (2021).
80. Ding, O. J. & Kennedy, G. J. Understanding vulnerability to late-life suicide. *Curr. Psychiatry Rep.* **23**, 58 (2021).
81. Raue, P. J., Ghesquiere, A. R. & Bruce, M. L. Suicide risk in primary care: identification and management in older adults. *Curr. Psychiatry Rep.* **16**, 466 (2014).
82. Becker, M. & Correll, C. U. Suicidality in childhood and adolescence. *Dtsch. Arztebl. Int* **117**, 261–267 (2020).
83. Bryan, C. J. & Rozek, D. C. Suicide prevention in the military: a mechanistic perspective. *Curr. Opin. Psychol.* **22**, 27–32 (2018).
84. Gonçalves, A., Sequeira, C., Duarte, J. & Freitas, P. Suicide ideation in higher education students: influence of social support. *Aten. Primaria* **46**, 88–91 (2014).
85. Bari, S. et al. The prevalence of psychotic symptoms, violent ideation, and disruptive behavior in a population with SARS-CoV-2 infection: preliminary study. *JMIR Form. Res.* **6**, e36444 (2022).
86. Vike, N. L. et al. The relationship between a history of high-risk and destructive behaviors and COVID-19 infection: preliminary study. *JMIR Form. Res.* **7**, e40821 (2023).
87. Lang, P., Bradley, M. & Cuthbert, B. *International Affective Picture System (IAPS): Affective Ratings of Pictures and Instruction Manual*. Technical Report A-8 (NIMH Center for the Study of Emotion and Attention, 2008).
88. Viswanathan, V. et al. A quantitative relationship between signal detection in attention and approach/avoidance behavior. *Front. Psychol.* **8**, 122 (2017).
89. Perlis, R. H. et al. Prevalence of incompletely penetrant Huntington's disease alleles among individuals with major depressive disorder. *Am. J. Psychiatry* **167**, 574–579 (2010).
90. Perlis, R. H. Association of a polymorphism near CREB1 with differential aversion processing in the insula of healthy participants. *Arch. Gen. Psychiatry* **65**, 882–892 (2008).
91. Gasic, G. P. et al. BDNF, relative preference and reward circuitry responses to emotional communication. *Am. J. Med. Genet. B Neuropsychiatr. Genet.* **150B**, 762–781 (2009).
92. Blood, A. J. et al. Microstructural abnormalities in subcortical reward circuitry of subjects with major depressive disorder. *PLoS ONE* **5**, e13945 (2010).
93. Makris, N. et al. Cortical thickness abnormalities in cocaine addiction—a reflection of both drug use and a pre-existing disposition to drug abuse? *Neuron* **60**, 174–188 (2008).
94. Viswanathan, V. et al. Age-related striatal BOLD changes without changes in behavioral loss aversion. *Front. Hum. Neurosci.* **9**, 176 (2015).
95. Posner, K. et al. The Columbia-suicide severity rating scale: initial validity and internal consistency findings from three multisite studies with adolescents and adults. *Am. J. Psychiatry* **168**, 1266–1277 (2011).
96. *Asking About Suicide is Vital for the Military and Veterans* (Columbia Lighthouse Project, 2016); https://cssrs.columbia.edu/the-columbia-scale-c-ssrs/military/
97. Livengood, S. L. et al. Keypress-based musical preference is both individual and lawful. *Front. Neurosci.* **11**, 136 (2017).
98. Shannon, C. E. & Weaver, W. *The Mathematical Theory of Communication* (Univ. Illinois Press, 1949).
99. Pedregosa, F. et al. Scikit-learn: machine learning in Python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).
100. Lemaitre, G., Nogueira, F. & Aridas, C. K. Imbalanced-learn: a Python toolbox to tackle the curse of imbalanced datasets in machine learning. *J. Mach. Learn. Res.* **18**, 1–5 (2017).
101. Hayes, A. F. Beyond Baron and Kenny: statistical mediation analysis in the new millennium. *Commun. Monogr.* **76**, 408–420 (2009).
102. MacKinnon, D. P., Fairchild, A. J. & Fritz, M. S. Mediation analysis. *Annu. Rev. Psychol.* **58**, 593–614 (2007).
103. Chen, Y. et al. Brain perfusion mediates the relationship between miRNA levels and postural control. *Cereb. Cortex Commun.* **1**, tgaa078 (2020).
104. Baron, R. M. & Kenny, D. A. The moderator-mediator variable distinction in social psychological research: conceptual, strategic and statistical considerations. *J. Pers. Soc. Psychol.* **51**, 1173–1182 (1986).
105. Markowitz, H. Portfolio selection. *J. Finance* **7**, 77–91 (1952).

Acknowledgements

Funding for this work was provided in part to H.C.B. by the Office of Naval Research (ONR award no. N00014-21-1-2216) and to H.C.B. (contact primary investigator) and A.K.K. (subcontract primary investigator) (ONR award no. N00014-23-1-2396) and to H.C.B. from a Jim Goetz donation to the University of Cincinnati (UC), College of Engineering and Applied Science (CEAS). We also thank C. Ross, A. Braggs-Brown, T. Talavage, E. Nauman and M. Cahay at UC CEAS, who greatly impacted the transfer of research funding to UC, allowing this work to be completed. The opinions expressed herein are those of

the authors and are not necessarily representative of those from their respective institutions.

Author contributions

The study was conceived and designed by H.C.B., A.K.K. and S.L. Acquisition of the original data was performed by H.C.B., B.-W.K., N.L.V., S.B., S.L., L.S., M.B. and A.K.K. Coding of statistical tools was carried out by S.L., S.B. and B.-W.K. (with guidance from H.C.B. and A.K.K.). Data were analyzed by S.L., S.B. and B.-W.K. (with guidance from H.C.B. and A.K.K.). Data interpretation was carried out by S.L., H.C.B. and A.K.K. (with input from B.-W.K., N.L.V., S.B., L.S., M.B. and N.M.). Statistical assessments were performed by S.L., S.B. and B.-W.K. (with guidance from H.C.B. and A.K.K.). The original draft was written by S.L. and H.C.B. Figures were generated by S.L., L.S., B.-W.K. and H.C.B. Revision of the manuscript for content was performed by all authors. All authors approved the final version of the paper for submission.

Competing interests

S.L., S.B., M.B., H.C.B., A.K., B.-W.K., L.S. and N.V. submitted a provisional patent "Systems and Methods Integrating Cognitive Science with Machine Learning for High Accuracy Prediction of Suicidal Thoughts and Behaviors". The provisional application is led by University of Cincinnati (Office of Innovation) in conjunction with Northwestern University, Application # 63/551,326. The other authors declare no competing interests.

Inclusion and ethics statement

De-identified data were collected by a third-party vendor (Gold Research) to reflect the general population demographics in the United States at the time of collection (December 2021) and with oversampling by 15% for mental health conditions. All participants provided informed consent, which included their primary participation in the study as well as the secondary use of their anonymized, de-identified data (that is, all identifying information removed by Gold Research Inc. before retrieval by the research group) in secondary analyses. Informed consent was obtained for all participants, as approved by the Institutional Review Board of Northwestern University (NU; approval no. STU00213665) for initial project start and later also approved by the University of Cincinnati (UC) Institutional Review Board (approval no. 2023-0164) as some NU investigators moved to UC. All work was done in accordance with the Declaration of Helsinki.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s44220-024-00229-x.

Correspondence and requests for materials should be addressed to Hans C. Breiter.

Peer review information *Nature Mental Health* thanks Katherine Schafer and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

[1]Department of Electrical and Computer Engineering, Northwestern University, Evanston, IL, USA. [2]Departments of Computer Science & Biomedical Engineering, University of Cincinnati, Cincinnati, OH, USA. [3]School of Medicine, Aristotle University of Thessaloniki, Thessaloniki, Greece. [4]Medill Integrated Marketing Communications, Northwestern University, Evanston, IL, USA. [5]Department of Radiology, Northwestern University, Chicago, IL, USA. [6]Department of Computer Science, Northwestern University, Evanston, IL, USA. [7]Laboratory of Neuroimaging and Genetics, Department of Psychiatry, Massachusetts General Hospital and Harvard School of Medicine, Boston, MA, USA. [8]These authors contributed equally: Sumra Bari, Nicole L. Vike, Leandros Stefanopoulos, Byoung-Woo Kim. [9]These authors jointly supervised this work: Aggelos K. Katsaggelos, Hans C. Breiter. ✉e-mail: breitehs@ucmail.uc.edu

nature portfolio

Corresponding author(s): Hans Breiter

Last updated by author(s): 2/29/24

Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our Editorial Policies and the Editorial Policy Checklist.

Please do not complete any field with "not applicable" or n/a. Refer to the help text for what text to use if an item is not relevant to your study.
For final submission: please carefully check your responses for accuracy; you will not be able to make changes later.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☒	☐	A description of all covariates tested
☐	☒	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☒	☐	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	Microsoft Excel (version 16.79.2) was used for tabulation of data.
Data analysis	Relative Preference variables were extracted using Matlab 7.1 with the following toolboxes: Curve Fitting Toolbox 1.1.4, Image Processing Toolbox 5.1, Matlab Builder for Excel 1.2.5, Statistics Toolbox 5.1 and Symbolic Math Toolbox 3.1.3. Machine learning analyses were implemented in Python 3.9 using the packages imblearn 0.0, sklearn 1.2.2, pandas 2. 0.2, and pandas 2. 0.2. Feature importance (Gini scores and mutual information scores) were obtained with sklearn 1.2.2. Figures relating to variable importance were plotted with seaborn 0.12.2. Mediation/Moderation analyses were conducted in R 4.2. 0 with libraries readxl 1. 4. 0, MASS 7. 3-56 and stats 4.2. 0.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio guidelines for submitting code & software for further information.

Data

Policy information about availability of data
All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

Anonymized data (relative preference variables, demographics and survey variables including suicidality variables) will be made publicly available on Open Science Framework pending acceptance of the manuscript. The data will be provided in Microsoft Excel.

Research involving human participants, their data, or biological material

Policy information about studies with human participants or human data. See also policy information about sex, gender (identity/presentation), and sexual orientation and race, ethnicity and racism.

Reporting on sex and gender

Gender was reported as assigned at birth. 38.09% of participants (1324) were male, 61.51 % of participants (2138) were female and 0.4% of participants (14) did not disclose their gender assigned at birth. 5.9% of participants (205) were between the ages of 18-24, 11.05% of participants (384) were between the ages of 25-34, 15.22% of participants (529) were between the ages of 35-44, 17.52% of participants (609) were between the ages of 45-54, 24.54% of participants (853) were between the ages of 55-64, and 25.78% of participants (896) were greater than 65 years of age.

Reporting on race, ethnicity, or other socially relevant groupings

2.91 % of participants (101) reported some high school education, 21.12% of participants (734) reported high school graduate education, 29.63% of participants (I 030) reported some college education, 22.35% of participants (777) reported bachelor's degree education, 5.06% of participants (I 76) reported some graduate school education, 16.23% of participants (564) reported graduate degree education and 2.7% of participants (94) reported post-doctoral training education. 83.66% of participants (2908) were White/Caucasian, 6.53% of participants (227) were African American, 3.54% of participants (123) were Hispanic/Latino, 3.51 % of participants (122) were Asian or Pacific Islander, 0.75% of participants (26) were Native American or Alaskan Native, 0.86% of participants (30) were Mixed racial background, 0.46% of participants (16) other socially relevant were another race and 0.69% did not disclose their race or ethnicity.86.02% of participants (2990) were right handed, 11. 71 % of participants (407) were left handed and 2.27% of participants (79) were ambidextrous.

Population characteristics

Suicidal thought and behavior was reported on a five-point Likert scale. For passive suicidal ideation 79 .17% of participants (2752) reported a score of 1, 6.88% of participants (239) reported a score of 2, 8.29% of participants (288) reported a score of 3, 3.60% of participants (125) reported a score of 4, and 2.07% of participants (72) reported a score of 5.For active suicidal ideation, 85.76% of participants (2981) reported a score of 1, 5.01 % of participants (174) reported a score of 2, 4.83% of participants (168) reported a score of 3, 2. 76% of participants (96) reported a score of 4 and 1.64% of participants (57) reported a score of 5.For planning for suicide, 88.55% of participants (3078) reported a score of 1, 3.77% of participants (131) reported a score of 2, 3.48% of participants (121) reported a score of 3, 2.7% of participants (94) reported a score of 4, and 1.50% of participants (52) reported a score of 5. For planning for safety, 82.8% of participants (2878) reported a score of 1, 3.62% of participants (126) reported a score of 2, 4.69% of participants (163) reported a score of 3, 3.68% of participants (128) reported a score of 4, and 5.21 % of participants (I 81) reported a score of 5.

5.90% of participants (205) were between the ages of 18-24, 11.05% of participants (384) were between the ages of 25-34, 15.22% of participants (529) were between the ages of 35-44, 17.52% of participants (609) were between the ages of 45-54, 24.54% of participants (853) were between the ages of 55-64 and 25.78% of participants (896) were at least 65 years of age. No participants under of the age of 18 were included in the cohort.

86.02% of participants (2990) were right handed, 11.71% (407) were left handed and 2.27% of participants (79) were ambidextrous.

Recruitment

Recruitment was conducted by a third party vendor, Gold Research (San Antonio, Texas) in December, 2021, who oversampled 15% of an initial recruitment pool for mental health conditions. Gold Research then surveyed subjects, assessed Recruitment for response completeness, and performed an initial quality screen of subjects. They released deidentified data for 4,019 subjects to the authors who performed a separate quality assurance check following published methods (see references 37,39,49,57) to leave 3,476 subjects for data analysis.

Potential self-selection bias relevant to the recruitment involved sampling individuals who used technology (a personal device was necessary to complete the survey and behavioral task) -- this bias may have confounded population characteristics.

Ethics oversight

All participants provided informed consent which included their primary participation in the study as well as the secondary use of their anonymized, de-identified data (i.e., all identifying information removed by Gold Research Inc. prior to retrieval by Ethics oversight the research group) in secondary analyses. Informed consent was obtained for all participants, as approved by the Institutional Review Board of Northwestern University (NU) (approval number STU00213665) for initial project start and later also approved by the University of Cincinnati (UC) Institutional Review Board (approval number 2023-0164) as some NU investigators moved to UC, to be in accordance with the Declaration of Helsinki.

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

□ Life sciences □ Behavioural & social sciences □ Ecological, evolutionary & environmental sciences

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size

A third-party vendor, Gold Research Inc. (San Antonio, Texas), recruited adults across the united states of ages 18-70. To ensure adequate samples of participants with mental health conditions, Gold Research oversampled 15% of the sample for mental health conditions. Participant demographics were matched to the U.S. Census Bureau at the time of sampling in December, 2021. 4019 adults decided to participate and completed all procedures; after applying data exclusion criteria, data from 3476 participants was retained. State-of-the-art machine learning algorithms were used to predict suicidal thoughts and behaviors. No sample size calculations were performed in advance as no standard methods exist to calculate the required sample size for machine learning models. We sought a sample size in the upper range of what has been published to date (please see Table S5), namely in the range of 4000 subjects, knowing there would be sample reduction from quality assurance. Such a sample would also be adequate for the number of partitions we were using with the cross-validation approach to prediction.

Data exclusions

Of the 4019 deidentified participants released to authors for data analysis, participants meeting at least one of the following six criteria were omitted from the cohort: (1) participants with ten or more clinician-diagnosed illnesses, (2) participants that selected the same response for at least one section of the survey, (3) participants that rated all images in the behavioral task the same or with a variance of one (meaning only two of seven Likert points were used), (4) participants whose relative preference analysis yielded extreme outliers > 3 IQRS or incomplete measurements, (5) participants that had mis-matching responses to years of education and education level in survey, (6) participants that completed the questionnaire in less than 800 seconds. Data exclusion reduced the sample from 4019 participants to 3476 participants for analysis.

Replication

The results were generated with 10-fold cross validation in Python to ensure reproducibility.

Randomization

Gold Research randomized particpant selection from their database that fell within the demographic criteria for ages 18-70. Machine learning fits were further also randomized using *sklearn* in *python* with 10-fold cross validation.

Blinding

Participants were anonymized by the third party vendor, Gold Research before being released to the authors for data for data analysis. Investigators were blinded to group allocation during data collection.

Behavioural & social sciences study design

All studies must disclose on these points even when the disclosure is negative.

Study description

Research sample

Sampling strategy

Data collection

Timing

Data exclusions

Non-participation

Randomization

Ecological, evolutionary & environmental sciences study design

All studies must disclose on these points even when the disclosure is negative.

Study description	
Research sample	
Sampling strategy	
Data collection	
Timing and spatial scale	
Data exclusions	
Reproducibility	
Randomization	
Blinding	

Did the study involve field work? □ Yes □ No

Field work, collection and transport

Field conditions	
Location	
Access & import/export	
Disturbance	

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	□ Antibodies
☒	□ Eukaryotic cell lines
☒	□ Palaeontology and archaeology
☒	□ Animals and other organisms
☒	□ Clinical data
☒	□ Dual use research of concern
☒	□ Plants

Methods

n/a	Involved in the study
☒	□ ChIP-seq
☒	□ Flow cytometry
☒	□ MRI-based neuroimaging

Antibodies

Antibodies used	
Validation	

Eukaryotic cell lines

Policy information about cell lines and Sex and Gender in Research

Cell line source(s)

Authentication

Mycoplasma contamination

Commonly misidentified lines
(See ICLAC register)

Palaeontology and Archaeology

Specimen provenance

Specimen deposition

Dating methods

□ Tick this box to confirm that the raw and calibrated dates are available in the paper or in Supplementary Information.

Ethics oversight

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Animals and other research organisms

Policy information about studies involving animals; ARRIVE guidelines recommended for reporting animal research, and Sex and Gender in Research

Laboratory animals

Wild animals

Reporting on sex

Field-collected samples

Ethics oversight

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Clinical data

Policy information about clinical studies

All manuscripts should comply with the ICMJE guidelines for publication of clinical research and a completed CONSORT checklist must be included with all submissions.

Clinical trial registration

Study protocol

Data collection

Outcomes

Dual use research of concern

Policy information about dual use research of concern

Hazards

Could the accidental, deliberate or reckless misuse of agents or technologies generated in the work, or the application of information presented in the manuscript, pose a threat to:

No | Yes

□ □ Public health
□ □ National security
□ □ Crops and/or livestock
□ □ Ecosystems
□ □ Any other significant area

Experiments of concern

Does the work involve any of these experiments of concern:

No | Yes

□ □ Demonstrate how to render a vaccine ineffective
□ □ Confer resistance to therapeutically useful antibiotics or antiviral agents
□ □ Enhance the virulence of a pathogen or render a nonpathogen virulent
□ □ Increase transmissibility of a pathogen
□ □ Alter the host range of a pathogen
□ □ Enable evasion of diagnostic/detection modalities
□ □ Enable the weaponization of a biological agent or toxin
□ □ Any other potentially harmful combination of experiments and agents

Plants

Seed stocks

Novel plant genotypes

Authentication

ChIP-seq

Data deposition

□ Confirm that both raw and final processed data have been deposited in a public database such as GEO.

□ Confirm that you have deposited or provided access to graph files (e.g. BED files) for the called peaks.

Data access links
May remain private before publication.

Files in database submission

Genome browser session
(e.g. UCSC)

Methodology

Replicates

Sequencing depth

Antibodies

Peak calling parameters

Data quality

Software

Flow Cytometry

Plots

Confirm that:

□ The axis labels state the marker and fluorochrome used (e.g. CD4-FITC).

□ The axis scales are clearly visible. Include numbers along axes only for bottom left plot of group (a 'group' is an analysis of identical markers).

□ All plots are contour plots with outliers or pseudocolor plots.

□ A numerical value for number of cells or percentage (with statistics) is provided.

Methodology

Sample preparation

Instrument

Software

Cell population abundance

Gating strategy

□ Tick this box to confirm that a figure exemplifying the gating strategy is provided in the Supplementary Information.

Magnetic resonance imaging

Experimental design

Design type

Design specifications

Behavioral performance measures

Imaging type(s)

Field strength

Sequence & imaging parameters

Area of acquisition

Diffusion MRI □ Used □ Not used

Preprocessing

Preprocessing software

Normalization

Normalization template

Noise and artifact removal

Volume censoring

Statistical modeling & inference

Model type and settings

Effect(s) tested

Specify type of analysis: □ Whole brain □ ROI-based □ Both

Statistic type for inference

(See Eklund et al. 2016)

Correction

Models & analysis

n/a	Involved in the study
□	□ Functional and/or effective connectivity
□	□ Graph analysis
□	□ Multivariate modeling or predictive analysis

Functional and/or effective connectivity

Graph analysis

Multivariate modeling and predictive analysis