Towards a Neurometric-based Construct Validity of Trust

Pin-Hao A. Chen[1,2,3], Dominic Fareri[4], Berna Güroğlu[5,6], Mauricio R. Delgado[7],& Luke J. Chang[8]

[1] Department of Psychology
[2] Neurobiology and Cognitive Science Center
[3] Center for Artificial Intelligence and Advanced Robotics
National Taiwan University
Taipei, Taiwan

[4] Gordon F. Derner School of Psychology
Adelphi University
Garden City, NY, 11530

[5] Institute of Psychology
Leiden University
Leiden, The Netherlands

[6] Leiden Institute for Brain and Cognition (LIBC)
Leiden, The Netherlands

[7] Department of Psychology
Rutgers University
Newark, NJ, 07102

[8] Department of Psychological and Brain Sciences
Dartmouth College
Hanover, NH, 03755

*Corresponding author: andyphchen@ntu.edu.tw & luke.j.chang@dartmouth.edu

Word count: 8547

Abstract

Trust is a nebulous construct central to successful cooperative exchanges and interpersonal relationships. In this study, we introduce a new approach to establishing construct validity of trust using "neurometrics". We develop a whole-brain multivariate pattern capable of classifying whether new participants will trust a relationship partner in the context of a cooperative interpersonal investment game (n=40) with 90% accuracy and find that it also generalizes to a variant of the same task collected in a different country with 82% accuracy (n=17). Moreover, we establish the convergent and discriminant validity by testing the pattern on eleven separate datasets (n=496) and find that trust is reliably related to beliefs of safety, inversely related to negative affect, but unrelated to reward, cognitive control, social perception, and self-referential processing. Together these results provide support for the notion that the psychological experience of trust contains elements of beliefs of reciprocation and fear of betrayal aversion. Contrary to our predictions, we found no evidence that trust is related to anticipated reward. This work demonstrates how "neurometrics" can be used to characterize the psychological processes associated with brain-based multivariate representations.

Introduction

The foundation of modern society is built upon our ability to successfully conduct cooperative social exchanges such as strategic coalitions, exchange markets, and systems of governance. Trust plays a central role in facilitating social exchange [1] based on its ability to reduce transaction costs and increase information sharing [2]. Successful interpersonal, business, and political transactions require trusting that a relationship partner will honor their agreement. Countries with formal institutions that protect property and contract rights are associated with higher perceptions of trust and civic cooperation, decreased rates of violent crime in neighborhoods [3], and increased economic growth [4]. From an interpersonal perspective, trust can be considered the psychological state of assuming mutual risk with a relationship partner to attain an interdependent goal in the face of competing temptations [5,6] which can be assayed using a two-person Investment Game [7,8]. In this game, a Trustor has the opportunity to invest a portion of a financial endowment to a Trustee. The investment amount is multiplied by a factor specified by the experimenter (e.g., 3 or 4), and the Trustee ultimately decides how much of the multiplied endowment to return back to the Trustor to honor or betray their trust. This game has been well studied in behavioral economics [9] and also in the field of decision neuroscience, which has investigated the neurobiological processes associated with trust [10–18] and its reciprocation [19,20]. This work has found that trust and reciprocity are associated with neural reward circuitry including the ventral striatum, ventral tegmental area (VTA), and medial prefrontal cortex. However, it remains unclear precisely how this neural circuitry produces psychological feelings of trust that drives behavior in interpersonal interactions. In this paper, we establish a "neurometric" approach to assessing the construct validity of brain activity patterns predictive of individual decisions to trust in the investment game.

Trust is a dynamic state that evolves over the course of a relationship. Early stages of a relationship are focused on assessing a partner's trustworthiness level, which can be impacted by previous interactions [13,14,18,21], gossip [22–24], group membership [25], or judgments based on appearance [26,27]. Trustors must be willing to endure a risk [28,29], while Trustees must be willing to overcome their own self-interest and take an action that fulfills an interdependent goal. As the relationship progresses, both parties are better able to predict each other's behavior and develop a sense of security in the relationship. In this way, trustworthiness reflects a dynamic belief about the likelihood of a relationship partner reciprocating [14,16,30–32]. These mutually beneficial collaborations can be rewarding [10,12,33]. However, at some point in the relationship, one person may end up betraying their partner [34], which could eventually lead to a dissolution of the relationship [17]. Thus, the candidate motivations influencing our likelihood to place trust in others include: (a) beliefs about probability of future reciprocation, (b) anticipated rewards, and (c) betrayal-aversion.

In psychometrics, creating a quantitative measurement of a nebulous and multifaceted concept such as trust requires establishing construct validity. Constructs provide consensus understanding of the semantic meaning of an abstract concept based on a nomological network of associations to other concepts [35]. Validating a construct requires assessing its generalizability to new populations and contexts and its convergent and discriminant validity to other constructs [36]. Though the principles of psychometrics were originally established for more traditional psychological tests and questionnaires, there is growing evidence that patterns of brain activity can serve as "neurometrics" of constructs [37]. For example, there has been a longstanding interest in using multivariate decoding methods to determine an individual's psychological state based on patterns of brain activity [38–41] with demonstrated success in predicting the intensity of a variety of affective experiences [42–46], reconstructing a visual stimulus [47] or uncovering its semantic meaning [48]. Neurometrics has several advantages over psychometrics in that it can utilize a high dimensional measurement of voxel activity observed during the engagement of a specific psychological process without requiring retrospective verbal self-report (e.g., questionnaires) or completing many different behavioral tasks (e.g., intelligence tests). By leveraging quickly changing scientific norms in open data sharing [49–53], it is increasingly possible to train a model predictive of a psychological state using brain activity such as pain [42], and establish a nomological network based on the model's convergent and discriminative validity with other constructs such as negative emotions [45], cognitive control [41], social rejection [54], and vicariously experienced pain [43].

Building on this approach, in this study we use supervised multivariate pattern-based analysis to predict individual decisions to trust a relationship partner in an interpersonal context using data from two previously published studies [12,14] (Figure 1A). We then establish the neurometric properties of this brain model by assessing its generalizability to a slightly different version of the task collected in a different country [55] (Figure 1B) and its convergent and divergent validity across 11 different tasks probing risk [56,57], affect [45,58], rewards [59–61], cognitive control [62], and social cognition [63–66]. This process allows us to characterize the psychological properties of the construct of trust using neurometric analyses (Figure 1C). Based on the findings outlined above, we hypothesize that the construct of trust will be positively associated with beliefs of safety, feelings of anticipated reward, and negatively with feelings of negative affect, but not associated with other psychological processes (Figure 1D).

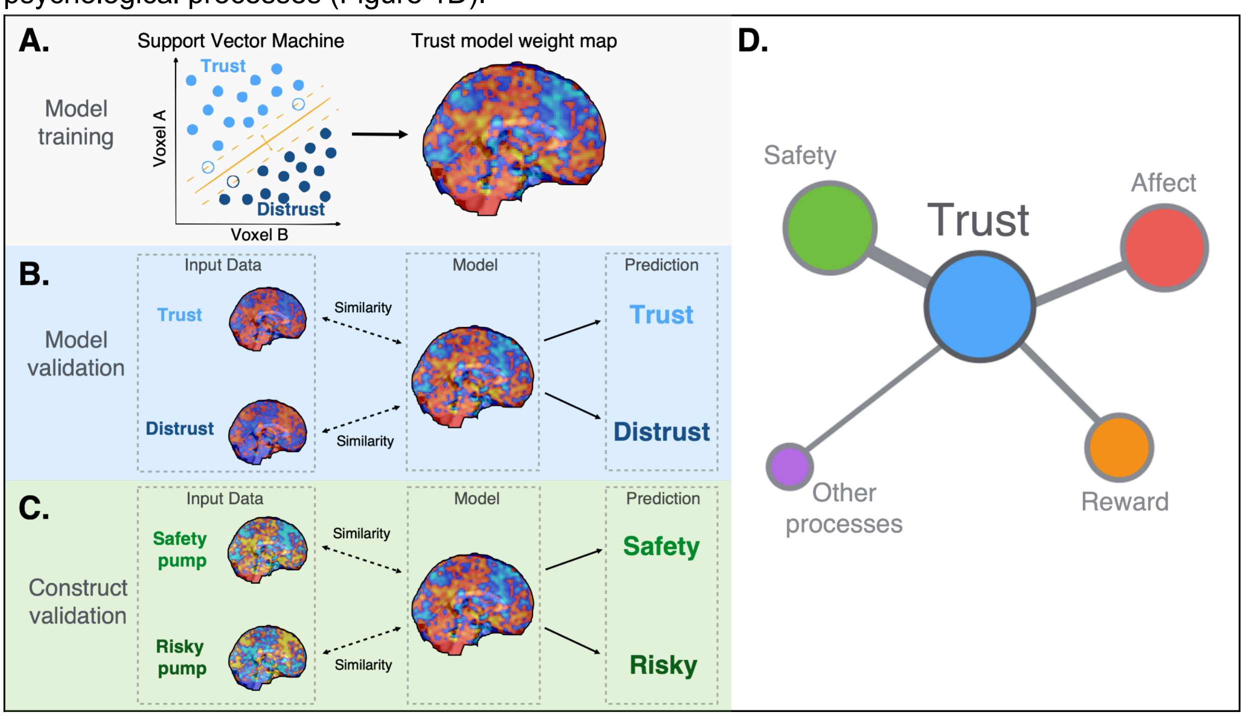

Figure 1. A demonstration of construct validity based on neurometric information. (A) A support vector machine algorithm was used to train the trust model. (B) An independent trust dataset was used to validate the trust model's generalizability. (C) We tested the model on independent datasets such as the Balloon Analog Risk Task to assess the convergent and discriminant validity of the trust model. (D) We hypothesized that trust was associated with beliefs of safety, feelings of anticipated reward, and affect, but not other processes.

Results

Training trust brain model

Using data from two published studies [12,14], we trained a linear Support Vector Machine (SVM) to classify when participants (n=40) decided to trust a relationship partner in the investment game using whole-brain patterns of brain activity (Figure 1A). We performed an initial temporal data reduction using univariate general linear models (GLMs) to create an average map of each participant's brain response when making decisions to trust or not. We then used a leave-one-subject-out (LOSO) cross-validation procedure to evaluate the performance of our multivariate SVM model in classifying maps associated with each participant's decisions to prospectively trust or distrust using data from the rest of the participants. Our trust brain model (Figure 2A) was able to accurately discriminate between trust and distrust decisions within each participant (forced-choice accuracy: 90%, $p < 0.001$, Figure 2B & 2C, Table S1). Forced choice tests compare the relative pattern expression of the model between brain maps within the same participant and are particularly well suited for fMRI because they do not require signals to be on the same scale across individuals or scanners [42].

To establish the face validity of our model, we used a parametric bootstrap to identify which voxels most reliably contributed to the classification, which involved retraining the model 5,000 times after randomly sampling participants with replacement. This procedure is purely for visualization and not used for spatial feature selection [67]. Consistent with prior work, we observed positive weights in the ventromedial prefrontal cortex (vmPFC), septal area [12,14,68], amygdala, and ventral hippocampus. Negative weights were found in the dorsal anterior cingulate cortex (dACC) and bilateral insula (Figure 2A). The pattern of weights learned across these bootstrap were highly reliable. We computed the pairwise spatial similarity of the whole brain pattern estimated across each bootstrap iteration and observed a high level of spatial consistency, r=0.91 [45].

Next, we trained a general trust model using data from all participants and evaluated its generalizability on a variant of the trust game in which participants receive feedback about their partner's decisions regardless if the participant decided to trust or not (Figure 1B). Importantly, we found that our model was able to accurately discriminate between the trust and distrust decisions from participants recruited from a different country collected on a different scanner (forced-choice accuracy: 82%, $p = 0.006$, Figure 2B, Table S1). This provides further confirmation that our model is capturing aspects of the psychological experience of trust that is shared across participants.

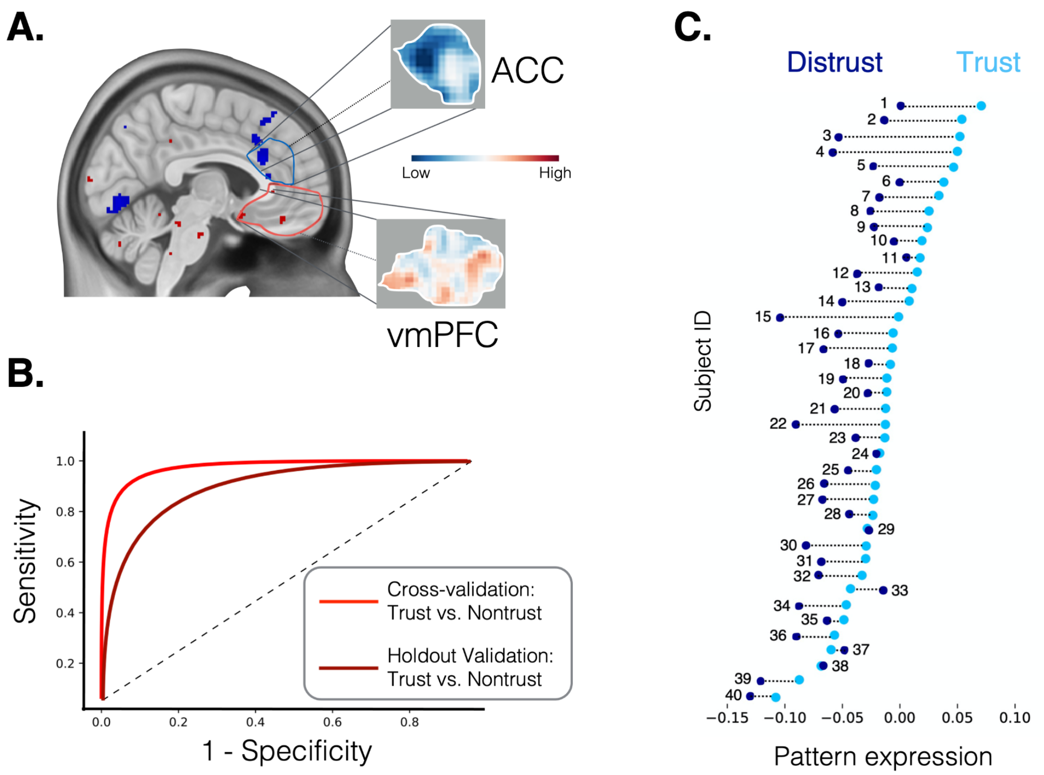

Figure 2. The trust model and its performance in the training and validation dataset. (A) The trust model is a whole brain pattern of voxel weights that can be linearly combined with new data to predict psychological levels of trust. We visualize the voxels that most reliably contribute to the classification using a bootstrap procedure (thresholded $p < 0.005$ uncorrected for visualization). (B) The receiver-operating-characteristic (ROC) plot highlights the sensitivity and specificity of the model in cross-validation and in an independent holdout dataset. (C) We plot the pattern expression, which reflects the spatial correlation between the model and decisions to trust and distrust across each of the 40 participants in the training dataset.

Construct Validity

After establishing the sensitivity of our model to accurately discriminate trust decisions, we next sought to evaluate the generalizability of the trust model to other psychological constructs using additional datasets. If the model performs at chance in other contexts, then this establishes the specificity of the model in capturing trust. However, if the model gets confused in other contexts, then this may reflect overlap in the psychological experience of trust to other related constructs.

Decisions to trust a relationship partner signal that the participant believes the partner is likely to reciprocate [30]. Trust reflects security in the relationship that the partner will behave as expected in their mutually interdependent interests. We first examined whether the trust model might be related to beliefs of safety, which can be measured using risk-taking tasks. The Balloon Analog Risk task (BART) is among the most widely used behavioral assay of risk-taking behavior [56,57]. In this task, participants are presented a series of colorful (the risk condition) or achromatic balloons (the safety or control condition) and are instructed to inflate the balloons. In the risk condition, participants can choose to inflate a balloon and only receive a reward if the balloon does not explode. However, each inflation is associated with an increasing probability of explosion, and when the balloon explodes, participants do not receive a reward for that round [56,57]. In contrast, in the safety condition, participants are also instructed to inflate a series of balloons, but there is no risk of the balloons exploding, nor an opportunity to receive a reward. We calculated the spatial similarity of our trust model to univariate beta maps from a GLM measuring average brain activity to the risk or safety conditions from two independent BART datasets (N=15 in dataset 3[56] and N = 123 in dataset 4[57]; Table S1). In both datasets, we found that the trust model could accurately discriminate between the safety and risk conditions (accuracy=93%, p < 0.001 in dataset 3; accuracy=93%, p < 0.001 in dataset 4; Figure 3B-2). These results indicate that the trust model captures a psychological experience that is shared with beliefs about safety when making risky choices (Figure 3A-2). When a relationship partner seems untrustworthy and reciprocation seems risky, participants will choose to keep their money rather than investing it.

Next, we explored if the trust model captured aspects of the experience related to negative affect. One reason why people may choose to distrust and not invest their money in a relationship partner is because of potential concerns about the partner betraying their trust and keeping all of the money. This results in negative utility for both losing money, and also being betrayed [34]. To test this hypothesis, we evaluated if the trust model might be inversely related to feelings of negative affect elicited by pictures from the international affective picture system (IAPS) from two independent datasets (Table S1). We found that in dataset 5 (N=93)[45], the trust model differentiated between conditions of neutral and negative emotional pictures (accuracy = 72%, p < 0.001; Figure 3B-3). A similar finding was also shown in dataset 6 (N=56) [58], where the trust model discriminated between the neutral and negative-valence picture conditions (accuracy = 69%, p = 0.002; Figure 3B-3) as well as between positive and negative-valence conditions (accuracy = 73%, p < 0.001; Figure 3B-3). These analyses provide evidence of overlap in the psychological processes associated with trust and negative affect. Specifically, decisions to trust are associated with less negative affect, consistent with a betrayal-aversion motivation. However, it is also possible that decisions to trust are associated with positive affect, but dataset 6 rules out this possibility. In this dataset, we did not observe a significant association with viewing positive compared to neutral pictures (accuracy = 50%, p = 0.551; Table S1), only positive and neutral compared to viewing negative pictures.

Third, we examined whether the trust model can be generalized to feelings of anticipated reward. We have previously demonstrated that learning that a close friend reciprocated is associated with a greater rewarding experience compared to when a stranger reciprocates [12], suggesting that trust may be associated with the anticipation of a future reward. To test this hypothesis, we evaluated whether our model was related to reward across three different tasks [59,60,69]. In dataset 7 (N=64; Table S1), participants guessed whether a randomly drawn card would be higher or lower than a specific number. If they were correct, they would receive a monetary reward, and if they were incorrect they would lose money [70]. We found that the trust model performed at chance in differentiating experienced rewards from losses [59] (accuracy = 42%, p = 0.921; Figure 3B-4). A similar result was also found in dataset 8 (N=18; Table S1), in which participants were shown either a cue indicating maximal gain or loss [60], and the trust model was unable to discriminate between these two conditions (accuracy = 66%, p = 0.119; Figure 3B-4). The trust model also failed to show a generalizability to discriminate anticipated rewards from no-rewards in the monetary incentive delay task [71] in dataset 9 (N=29; Table S1; accuracy = 38%, p = 0.933; Figure 3B-4) [61]. Thus, contrary to our hypotheses, our results revealed that the trust model has no clear association with feelings of anticipated reward across all three datasets related to anticipated or experienced rewards (Figure 3A-2).

Specificity of trust model

There are many other potential psychological aspects of the trust experience that can be evaluated using this neurometric approach. First, it is possible that people may vary in their preferences for selfishness and cooperation, and choosing to trust may involve overriding selfish motivations, which would require exhibiting cognitive control [72,73]. We tested this hypothesis by applying the model to a stop signal task (dataset 10; N=19; Table S1) [62], in which participants are instructed to override a prepotent response, and found that the trust model was unable to discriminate between the successful inhibition and inhibition failure conditions (accuracy = 57%, p = 0.326; Figure 3B-5). Decisions to trust may also require social cognition to consider the other player's mental states such as their beliefs, preferences, and financial outcomes. In order to demonstrate the specificity of the trust construct, we additionally tested our model on several datasets probing distinct aspects of social cognition. We found that the trust model did not generalize to perceptual judgments such as familiarity, in which participants judged whether a face is familiar or unfamiliar to the participants (dataset 11; N=16; accuracy = 63%, p = 0.217; Figure 3B-5; Table S1) [63]. We also found that the trust model did not generalize to the classification between viewing social and non-social scenes in dataset 12 (N=36; accuracy = 47%, p = 0.686; Figure 3B-5; Table S1) [64]. Lastly, we tested if the trust model was similar to self-referential cognition in a task in which participants made self-referential judgments or perceptual judgments (e.g., type of font) to a variety of trait adjectives (dataset 13; N=27; accuracy = 40%, p = 0.876; Figure (B-5); Table S1) [65,66]. Together, these findings indicate that the trust model was not associated with cognitive control, social perception, or self-referential processing (Figure 3A-2).

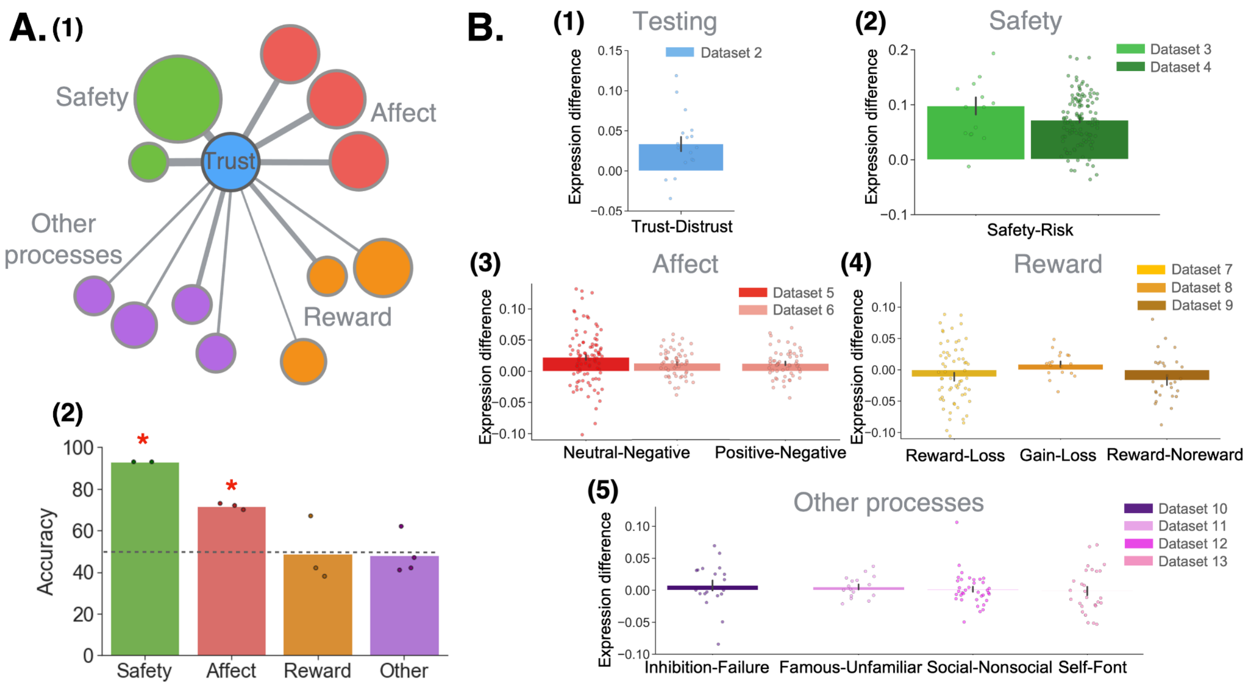

Figure 3. Construct validity of the trust model and model generalizability. (A) (1) Network plot illustrates that the trust model significantly generalizes to safety and affect datasets, but not to reward and other processing datasets. The distances and thickness of edges are weighted based on every 10% decrease in classification accuracy, and the size of nodes represents sample size of each dataset. (2) The forced-choice classification accuracy for each dataset within the four domains was shown in the bar plot. Only the safety and affect domains demonstrated above chance accuracy across datasets. (B) Trust model pattern expression differences between the two conditions in the: (1) trust testing datasets, (2) two safety datasets, (3) two affect datasets, (4) three reward datasets, as well as (5) four datasets involving cognitive control and social cognition.

Trust Nomological Network

Finally, we constructed a nomological network of psychological states by computing the spatial similarity of patterns of brain activity elicited by different experimental tasks. To do so, we first used Uniform Manifold Approximation and Projection (UMAP) [74], a nonlinear dimensionality reduction technique to visualize similarities of whole brain spatial patterns across all participants (N=553) from all thirteen datasets. We found that whole-brain multivariate patterns of trust were closer to those of beliefs of safety, feelings of anticipated no-reward or loss, and feelings of neutral or positive affect (Figure 4A). By contrast, whole-brain multivariate patterns of distrust were closer to those of beliefs of risk, feelings of anticipated reward, and feelings of negative affect (Figure 4A). Similar findings were also revealed in several brain regions, such as vmPFC, dmPFC and dACC (Figure S1).

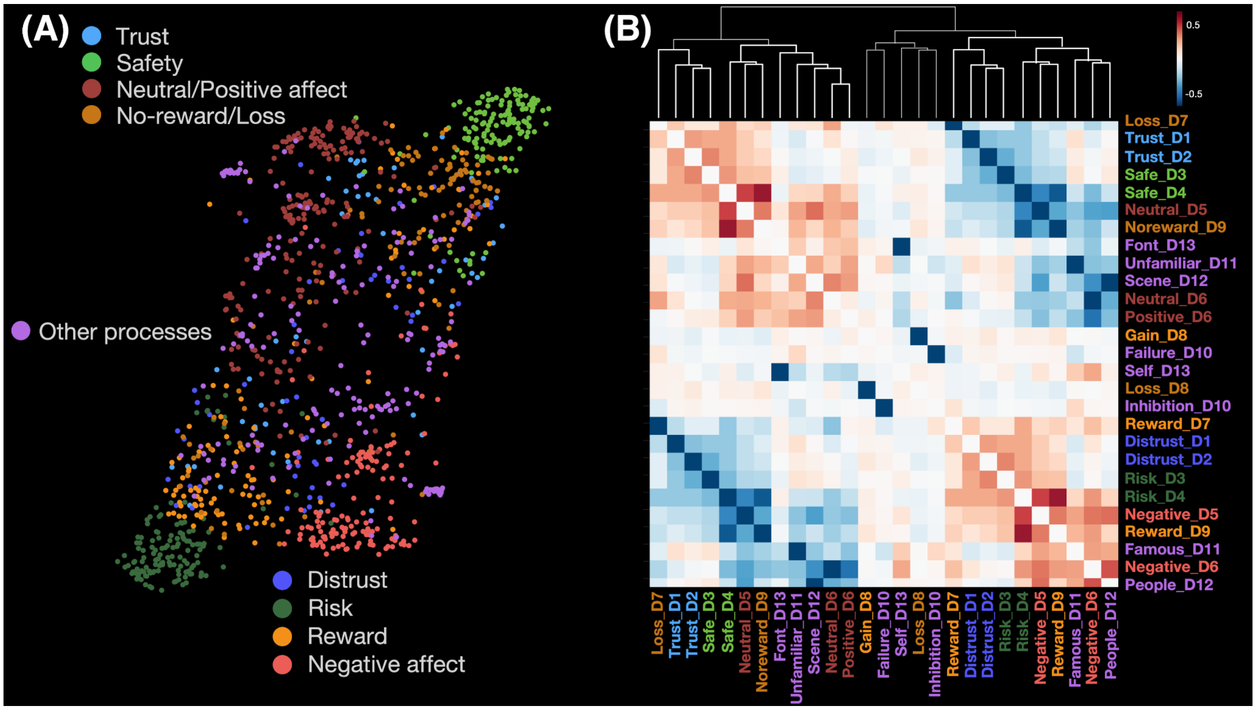

Figure 4. Spatial pattern similarity across all brain data from thirteen datasets. (A) Based on whole-brain spatial patterns, trust was more similar to safety, no-reward, neutral and positive affect; whereas distrust was more similar to risk, reward, and negative affect. (Each dot represents a beta map from each participant) (B) Hierarchical clustered heatmap of correlation across the mean spatial pattern from each condition (26 conditions from 13 datasets) also revealed similar findings as above.

We quantitatively verified the results from the UMAP visualization by computing the average brain response across participants for each condition of each task and then evaluating the pairwise spatial similarity of these maps (Figure 4B). We found that the spatial patterns between two trust conditions were highly similar to each other ($r = 0.33$), the training trust condition in dataset 1 was similar to the two safety conditions ($r = 0.31$ for dataset 3 and $r = 0.30$ for dataset 4, respectively), and the validation trust condition in dataset 2 was also similar to the two safety conditions ($r = 0.39$ for dataset 3 and $r = 0.30$ for dataset 4, respectively). In contrast, the spatial patterns between the two distrust conditions were highly similar to each other ($r = 0.32$), the training distrust condition in dataset 1 was similar to the two risk conditions ($r = 0.31$ for dataset 3 and $r = 0.30$ for dataset 4, respectively), and the validation distrust condition in dataset 2 was also similar to the two risk conditions ($r = 0.39$ for dataset 3 and $r = 0.30$ for dataset 4, respectively). In addition, the safety condition in dataset 4 revealed similar patterns to neutral/positive emotion conditions ($r = 0.30$ for neutral emotion and $r = 0.19$ for positive emotion in dataset 6), and the risk condition in dataset 4 was similar to negative emotion conditions ($r = 0.52$ for dataset 5 and $r = 0.34$ for dataset 6, respectively; Figure 4B). This means that, based on brain spatial patterns, the construct of trust is more closely related to beliefs of safety, anticipated no-reward, and non-negative affect, but not to beliefs of risk, anticipated reward, or negative affect.

Discussion

In this study, we sought to create a model of trust based on patterns of brain activity elicited during an interpersonal investment task. We employed a neurometric approach [37,40,42,43,45] to characterize this model by assessing its reliability and validity using multiple previously published open datasets. This model leverages reliable patterns of brain activity and is sensitive to detecting psychological states of trust that generalizes to new subjects, scanners, and variants of the investment game task. In addition, we also assessed the validity of our model [36]. Prior work has primarily relied on establishing face validity by demonstrating that regions associated with a construct have a reliable independent contribution to the prediction [42,43,45]. However, directly interpreting the weights of linear models can potentially be misleading [75]. An alternative approach based on the principles of construct validity attempts to triangulate a construct by establishing convergent and discriminant validity with respect to related and distinct constructs probed using multiple methods [36,41]. This has also been described as establishing a "nomological network" [35] and identifying the "receptive field" of a model [37]. We assessed the ability of our trust model to discriminate task conditions across a variety of potentially related psychological constructs elicited using many different types of tasks across 11 previously published datasets.

Overall, we found that our brain model of trust was associated with a distinct signature of related psychological processes. First, previous work has established that trust reflects dynamic beliefs about the likelihood of a relationship partner overcoming self-interest and reciprocating [13,30]. We find strong evidence supporting this interpretation. Across two separate experiments exploring risky decision-making, our trust model is reliably associated with safety compared to risk, or in other words, a high degree of certainty in avoiding a negative outcome compared to more uncertainty in the risky condition. In addition, our pattern similarity analyses indicate that decisions to not trust are associated with the risky conditions, while the trust conditions are associated with the safety conditions. Second, we find support for the hypothesis that trust requires overcoming concerns of potential betrayal [17,34]. We find that our trust model is reliably negatively associated with the psychological experience elicited from viewing negative arousing images relative to viewing neutral or positive images. We did not observe a significant relationship with differences between positive vs neutral indicating that it is neither positive nor neutral images driving this effect. Moreover, pattern similarity analyses revealed that viewing negative images correlated with the risky decision condition, while the neutral images correlated with the safety decisions. These findings are consistent with a betrayal-aversion account. It has been hypothesized that people may choose to keep their money and avoid investing in a relationship partner not just because they don't want to lose their money, but also because they want to avoid feeling betrayed by another person [34]. Of course, viewing negative arousing images is hardly the same thing as being betrayed and we believe this finding should be further substantiated in future work. Third, contrary to our predictions, we found no evidence that trust is associated with experiencing or anticipating a future reward. We tested our trust model on 3 distinct tasks probing the anticipation and experience of reward and found no indication that trust was related to reward or its anticipation. We think this is particularly important as it has been often assumed that the main motivation for trusting a relationship partner in the trust game is because the expected value is higher [9,12,30]. Our findings suggest that it is not the reward, but rather the probability calculus that may be driving decisions to trust. Finally, we also find that trust does not appear to be related to overcoming a prepotent tendency to be selfish, which would recruit cognitive control. Nor does it appear to be involved in social perceptual judgments such as whether an image is a person or an object, or if a person has been seen before or is new. We also find no evidence suggesting that trust involves self-referential processing such as considering self-other relative payoffs [76].

There are several important considerations when interpreting our results. First, we made no assumptions about potential brain regions that may be involved in the psychological experience of trust and chose to utilize a whole-brain approach when training our model [67]. This demonstrates which regions independently and additively contribute to the prediction. However, it is highly likely that brain activity may be highly collinear, which may lead to instability of the model weights [75,77]. We used a bootstrap approach to iteratively retrain the model using different subsets of the data and found that the regions with the largest weights were highly consistent (r=0.91). Future work may consider additionally exploring different types of spatial feature selection [78]. Second, our model is currently ignoring interactions between brain regions, which may be an important signature of the trust construct. This might be explored in the future by training new models using functional connectivity or interactions between brain regions. Third, our model is also agnostic to individual differences. We have established that the model generalizes to new participants, but it is not currently able to assess variations in potential motivations (e.g., risk-aversion vs betrayal-aversion). Future work might use multivariate methods for probing individual differences such as intersubject representational similarity analysis [20,79,80]. Finally, our construct validity analyses are completely dependent on the reliability and validity of the additional tasks, which has never really been fully established. In addition, we have only tested our model on a subset of the possible related constructs. We see this as an iterative process that cannot be fully addressed by a single paper, but instead will require continued refinement as more datasets become available in the future [37].

In summary, using 14 datasets, we establish a neurometric-based construct validity of trust. This model is stored as a three-dimensional brain image that contains a recipe for how to linearly combine information from each voxel in the brain[81]. Importantly, this model generalizes beyond the specific subjects, scanner, or experimental paradigm and can easily be shared with other researchers [37]. In addition, we move beyond a reverse inference approach [82] in interpreting the psychological processes associated with trust based on which regions contribute to the prediction [40,45], to a more quantitative construct validity approach. These analyses support several previous accounts of trust, but importantly rule out a reward-based motivation. This provides a proof of concept that brain activity can be used to make inferences about a psychological process beyond self-report or behavioral observations. We believe this general approach could be applied to any other psychological constructs that can be measured using patterns of brain activity.

Methods

fMRI Dataset

Trust model training datasets. The training datasets (dataset 1-1 and 1-2) for the trust model contained data from two published studies [12,14]. In dataset 1-1, 17 participants played an iterated trust game with three different trustees while undergoing fMRI in a 3T Siemens Allegra scanner (TR=2000ms; TE=25ms)[14]. Participants were endowed with one dollar and on each trial decided whether to invest this money in the other trustee (i.e., trust) or keep it (i.e., distrust). Decisions to trust resulted in the one dollar investment being multiplied by a factor of three. The trustee then decided whether to keep all three dollars, or share half of the return on the investment back to the participant (i.e., $1.50). In dataset 1-2, 23 participants also played a similar iterated trust game with two different trustees while undergoing fMRI in a 3T Siemens Magnetom Trio scanner (TR=2000ms; TE=30ms) [12].

In total, there were 40 participants from dataset 1-1 and 1-2 in the current study. We focused our analysis only on the decision epoch when participants made decisions to either trust or distrust. fMRI data were analyzed using a combination of custom scripts (https://github.com/rordenlab/spmScripts) for SPM12 and FSL (v5.09; FMRIB). We performed standard preprocessing in SPM (motion correction, brain extraction and coregistration, slice time correction). Motion artifact was removed using ICA-AROMA in FSL (Pruim et al., 2015). Functional data were smoothed using a 5mm kernel in FSL. Each condition was modeled as a separate regressor in a general linear model (GLM). This included a regressor modeling each of the decision types (trust or distrust) and the different possible decision outcomes (though these data were not the focus of the present manuscript). The GLM resulted in a trust whole-brain beta map and a distrust whole-brain beta map for each trustee (detailed preprocessing and GLM steps see [12,14]). We then averaged the beta maps across partner types within each participant. These maps were mean-centered values across all voxels within each beta map [83] and used to train the trust model.

Trust model validation dataset. The validation dataset (dataset 2) contained data from 17 participants (mean age = 20.6 years, SD=1.49; 24% female) who participated in a repeated trust game while undergoing fMRI in a 3T Philips Achieva scanner (TR = 2200 ms, TE = 30 ms, FOV = 220 × 220 × 114.7 mm; see [55] for more details about the sample and scanning parameters). All participants provided informed consent and the study was approved by the institutional review board at Leiden University Medical Center. Participants were instructed to play a trust game with three different targets, including a friend, an antagonist, and an anonymous peer. The game was designed to be slightly more similar to a prisoner's dilemma in that both players made their decisions simultaneously. Unlike a traditional trust game, participants received information about their partner's decisions regardless if they chose to share or keep. However, the responses from these targets were pre-determined by the computer and not the actual partner. Similar to dataset 1, we also focused our analysis on the decision epoch when participants made either a trust or distrust decision. Image pre-processing and analysis was conducted using SPM8 software (www.fil.ion.ucl.ac.uk/spm) implemented in MATLAB R2010 (MathWorks). Pre-processing included slice-time correction, realignment, spatial normalization to EPI templates, and smoothing with a Gaussian filter of 8 mm full-width at half maximum (FWHM). The fMRI time series were modeled by a series of events convolved with a canonical hemodynamic response function (HRF). The data was modeled at choice and feedback onset as null duration events. During decision-making the choice events (i.e., trust and keep decisions) were modeled for each of the three partner types. These modeled events were used as regressors in a general linear model (GLM) with a high pass filter using a discrete cosine basis set with a cutoff of 120 seconds. The GLM resulted in a trust whole-brain beta map and a distrust whole-brain beta map for each target. We then computed the mean trust whole-brain beta map across all three targets and repeated the same procedure for computing the mean distrust whole-brain beta map within each participant. We then mean-centered values across all voxels within each of the beta maps for all participants, and the mean-centered beta maps were used as a novel trust dataset for brain model validation.

Safety datasets. In order to test whether trust is associated with beliefs of safety, we had two open fMRI datasets using the Balloon Analog Risk Task (BART), in which one condition probes beliefs of risk and another probes beliefs of safety in this study. The BART aims to elicit naturalistic risk-taking behaviors, and each participant received two conditions in the fMRI scanner. In the risk condition, each inflation of balloons is a risky choice (pump), whereas inflating balloons in the safe condition is not a risky choice (control pump). In dataset 3 (OpenfMRI ds000001) [56], there are 15 healthy participants who underwent the two conditions in a 3T Siemens Allegra MRI scanner. The data were preprocessed by FSL (www.fmrib.ox.ac.uk/fsl), including realignment, highpass-filtering, brain extraction with BET, motion correction, spatial normalization, and smoothing with a 5 mm FWHM Gaussian kernel. For trials in the risky condition, the risky inflation and the other two task-related regressors were modeled separately in the GLM. For trials in the safe condition, the safe inflation and the other two task-related regressors were also modeled separately in the GLM. For each participant, the GLMs resulted in a risk inflation whole-brain beta map and a safe inflation whole-brain beta map. We then mean-centered values across all voxels within each beta map for all participants, and these mean-centered beta maps were used as data representing the risk condition and safety condition in the generalization testing.

In dataset 4 (OpenfMRI ds000030) [57], there are 123 healthy participants who also underwent the two conditions in a 3T Siemens Allegra MRI scanner. The data were preprocessed by FMRIPREP version 0.4.4, including motion correction, skullstripping and coregistration to T1 weighted volume, applying brain masks, realignment, normalization, and spatial smoothing with a 5 mm FWHM Gaussian kernel [84]. A risk inflation (accept pump) and a safe inflation (control pump), along with the other seven regressors were modeled in the GLM. For each participant, the GLM resulted in a risk inflation whole-brain beta map and a safe inflation whole-brain beta map, and we then mean-centered values across all voxels within each beta map for all participants. These mean-centered risk inflation beta maps and safe inflation beta maps were then used as data representing the risk condition and safety condition in the generalization testing.

Affect datasets. Two affect datasets were included in the current study. Dataset 5 came from the PINES dataset [45], which was an open dataset on Neurovault (https://identifiers.org/neurovault.collection:503). In this dataset, participants were asked to view numerous negative and neutral-valenced pictures from the international affective picture system (IAPS), and then rated how negative they felt from 1 (neutral) to 5 (most negative). Details of the experimental design were described in previous studies [45,85]. Among these participants, this current study only used data from those (N = 93) whose ratings had 1 (neutral) and 5 (most negative). The fMRI data was collected in a Siemens Trio 3T scanner (TR= 2000 ms, TE=29ms), and then preprocessed by SPM8, including unwarping, realignment, coregistration, normalization, spatial smoothing with a 6 mm FWHM Gaussian kernel and high pass filtering (180 sec cutoff). Then five separate regressors indicating different rating levels (1 to 5) were modeled in the GLM for each participant as well as 24 covariate regressors modeled movement effects (6 realignment parameters demeaned, their 1st derivatives, and the squares of these 12 regressors). Since our goal was to compare the neutral and negative condition, only the neutral (rating = 1) beta map and the negative (rating = 5) beta map were included in the current study. We then mean-centered values across all voxels within each of the above two kinds of beta maps for all participants. These mean-centered neutral and negative beta maps were taken as data in the generalization testing.

In Dataset 6, fifty-six participants were recruited to complete an emotional scene task [58]. In this task, participants were asked to make indoor/outdoor categorization judgments on scenes in a block design. Each block lasted 15 seconds and consisted of six emotional scenes with the same emotional valence. Each emotional-scene block alternated with a 15-sec fixation block, and each participant went through five blocks for each of three different valences, including positive, neutral, and negative valence. The emotional valence of the scenes used in each condition were selected from the IAPS and have been validated in a previous fMRI study [86]. The fMRI data was collected in a Philips Intera Achieva 3T scanner (TR = 2500 ms, TE = 35 ms), and then preprocessed by SPM8, including slice timing correction, unwarping, realignment, coregistration, normalization, and spatial smoothing with a 6 mm FWHM Gaussian kernel. The positive, neutral, and negative valence conditions were then modeled separately in the GLM for each participant. The GLM resulted in a positive, neutral, and negative emotion beta map from each participant, and we then mean-centered values across all voxels within each beta map for all participants. These mean-centered beta maps were used in the current study, representing three different emotional-valence conditions in the generalization testing.

Reward datasets. Three reward anticipation fMRI datasets were used in the current study. Dataset 7 comes from the Human Connectome Project [59], and the reward anticipation task used in this dataset is the Card Gambling task [70]. In this task, participants were asked to guess whether the number on a mystery card is greater or smaller than five. Participants would receive a reward of one dollar if the number is greater than five; by contrast, they would lose fifty cents if the number is smaller than five. In total, fMRI data from 64 participants were collected and preprocessed with the HCP fMRIVolume pipeline [87]. The preprocessing steps included gradient unwarping, motion correction, fieldmap-based EPI distortion correction, coregistration, normalization, and spatial smoothing with a 4 mm FWHM Gaussian kernel. The reward and loss conditions were then modeled in the GLM. The GLM resulted in a reward beta map and a loss beta map within each participant, and we then mean-centered values across all voxels within each beta map for all participants. These mean-centered reward and loss beta maps were then used as data representing the reward condition and non-reward/loss condition in the generalization testing.

In dataset 8, eighteen participants completed a reward/loss anticipation task while undergoing scanning in a 3T Siemens Trio scanner [60]. In this task, different cues were shown on the screen indicating different amounts of monetary reward or loss. After the cue phase, an outcome phase occurred, indicating the actual amount of reward and loss. The monetary reward or loss amounts were equally sampled from [1, 5, 20, 100]. The data were preprocessed using BrainVoyager QX 2.8 and NeuroElf V1.1, including: motion correction, slice timing correction, high-pass filtering, normalization, and spatial smoothing with a 6 mm FWHM Gaussian kernel. The cue and outcome phases with different levels were modeled separately in the GLM for each participant. Only the maximal-reward (i.e., a gain of $100) and maximal-loss (i.e., a loss of $100) beta maps from each participant were used in the current study, and we then mean-centered values across all voxels within each beta map for all participants. These mean-centered beta maps would represent the reward condition and non-reward/loss condition in the generalization testing.

In dataset 9 (OpenNeuro ds003242) [61], twenty-nine participants underwent a monetary incentive delay task [71,88] in an fMRI scanner. Before the task, participants were asked to memorize five abstract art images, and these familiar images were then taken as cues in the reward condition. In the reward condition, after a familiar image was shown on the screen as a reward cue, a number ranging from 1 to 9 was shown and participants had to respond whether the number was larger or smaller than 5. If participants responded fast enough (< 500 ms), they would receive a reward of one dollar. In the other condition, the non-reward condition, after a new abstract art image was shown as a non-reward cue, a number was also shown on the screen and participants were also asked to respond whether the number is greater or smaller than 5. However, the responding performance in the non-reward condition was not associated with any reward. FMRIPREP [89] was used for brain data preprocessing, and the steps included motion correction, skullstipping and coregistration to T1 weighted volume, applying brain masks, realignment, normalization, and spatial smoothing with a 6 mm FWHM Gaussian kernel. We modeled the reward condition and non-reward conditions as separate regressors in a univariate GLM, along with 24 covariate regressors modeling movement effects (6 realignment parameters demeaned, their 1st derivatives, and the squares of these 12 regressors), a 128 sec high pass filter using a discrete cosine transform, and separate scanner spikes based on frame differences that exceeded 3 standard deviations. For each participant, the GLM resulted in a reward beta map and a non-reward beta map, which were then mean-centered across all voxels within each beta map for all participants. These mean-centered reward and non-reward beta maps were then used as data representing the reward condition and non-reward/loss condition in the generalization testing.

Other processing datasets. In order to demonstrate the specificity of our trust model, we validated our model on four additional datasets, including cognitive control, familiarity, social cognition and self-referential cognition. To test the domain of cognitive control, in Dataset 10, nineteen participants performed a stop-signal task (SST) in a 3T Siemens Allegra MRI scanner (TR=2000ms, TE=30ms) [62]. This open dataset is available on both OpenNeuro (ds000007) and Neurovault (https://neurovault.org/collections/1807/). We used data from Neurovault task001, which was a manual SST. For the go trials in this task, participants were asked to press on the right or left button according to whether the letter "T" or "D" was shown on the screen. For stop trials, an auditory tone cue signaling stop was played after the letter being shown with some delay (stop-signal delay; SSD), and participants were asked to inhibit their approaching responses toward the button. Throughout the task, the length of SSD changed according to whether participants succeeded or failed to inhibit their responses in order to maintain the accuracy rate at 50%. Thus, the number of inhibition-success and inhibition-failure trials would be the same, and we would use data from both of these two conditions for further analysis. The data was preprocessed by FSL version 3.3, and the preprocessing steps included coregistration, realignment, motion correction, denoising using MELODIC, normalization, spatial smoothing with a 5 mm FWHM Gaussian kernel, and high-pass filtering. Details about the preprocessing steps were described in the original study [62]. Four conditions, including go, inhibition-success, inhibition-failure, and nuisance events, were modeled separately in the GLM for each participant. The GLM resulted in a go, inhibition-success, inhibition-failure, and nuisance event beta map, and we then mean-centered values across all voxels within each beta map for all participants. Only the mean-centered inhibition-success and inhibition-failure beta maps were used in the generalization testing.

For the domain of familiarity, in Dataset 11, sixteen participants completed a face-viewing task in a Siemens 3T TRIO scanner (TR=2000ms, TE=30ms) [63]. This open dataset is available on both OpenNeuro (ds000117) and Neurovault (https://neurovault.org/collections/1811/), and we used data downloaded from Neurovault. In this face-viewing task, participants were asked to view three different types of faces, including famous, unfamiliar, and scrambled faces. Each trial began with a fixation cross on the screen, and then one of the three types of faces were shown on the screen. Participants were asked to pay attention to all trials throughout the whole experiment. The fMRI data was preprocessed by SPM8, which included slice timing correction, realignment, coregistration, normalization, and spatial smoothing with a 8 mm FWHM Gaussian kernel. Three conditions, including famous, non-familiar, and scrambled faces were modeled separately in the GLM for each participant. The GLM resulted in a famous, non-familiar, and scrambled beta map, and we then mean-centered values across all voxels within each beta map for all participants. Only the mean-centered famous and non-familiar beta maps were used in the current study for the generalization testing.

For the domain of social cognition, in Dataset 12, thirty-six participants completed a scene judgment task in a Philips Intera Achieva 3T scanner (TR = 2500 ms, TE = 35 ms) [64]. In this task, each participant was asked to make indoor/outdoor categorization judgements on 270 different scenes, including 90 social scenes, 90 non-social scenes, and another 90 food scenes. These pictures have been used in several studies [90–92], and compared to non-social scenes, social scenes have been found to reliably activate brain regions, such as the dmPFC, PCC, and vmPFC [90]. In each trial, a scene image was shown on the screen for 2000 ms, followed by a 500 ms fixation, and a jitter (range: 0-5000 ms) was followed between each trial. The fMRI data were preprocessed by SPM8, which included slice timing correction, unwarping, realignment, motion correction, normalization, spatial smoothing with a 6 mm FWHM Gaussian kernel. The social, non-social, and food conditions were then modeled separately in the GLM for each participant. The GLM resulted in a social, non-social, and food beta map, and we then mean-centered values across all voxels within each beta map for all participants. Only the mean-centered social and non-social beta maps from each participant were then used in the current study, representing the social and non-social condition for the generalization testing.

For the domain of self-referential cognition, in Dataset 13, twenty-seven participants completed a trait-judgment task in a Philips Intera Achieva 3T scanner (TR = 2500 ms, TE = 35 ms) [65,66]. In this trait-judgment task, participants were asked to make three different targets of judgments, including self-judgment (i.e. does this adjective describe you?), mother-judgment (i.e. does this adjective describe your mother?) and font-judgment (i.e. is this adjective printed in bold-faced letters?) in two different languages. For each trial, a trait adjective word (e.g. smart) was paired with a target word (i.e. SELF, MOTHER, or FONT) and were shown on the screen for 2500ms. Although each trait word was presented once in Mandarin and once in English, the current study only used the three conditions in Mandarin. The fMRI data was preprocessed by SPM8, which included slice timing correction, unwarping, realignment, motion correction, normalization, spatial smoothing with a 6 mm FWHM Gaussian kernel. The self-judgment, mother-judgment, and font-judgment conditions were then modeled separately in the GLM for each participant. The GLM resulted in a self-judgment, mother-judgment, and font-judgment beta map, and lastly we mean-centered values across all voxels within each beta map for all participants. Only the mean-centered self-judgment and font-judgment beta mps from each participant were then used for the generalization testing in the current study.

Training and validating a trust model

Training model and cross-validation within the training dataset. We used a three stage approach to train our whole-brain multivariate classification model using a linear Support Vector Machine (SVM). First, we were interested in evaluating how well the model might generalize to new data using a leave-one-subject-out (LOSO) cross-validation procedure, ensuring that every subject served as both training and testing data [45]. This allowed us to evaluate how a model trained on 39 participants could classify trust or distrust decisions from the left-out participant and provided an estimate of the expected generalizability of the model to similar datasets. Second, we were interested in assessing which voxels most reliably contributed to the trust classification. We used a parametric bootstrap procedure, which involved retraining the model 5,000 times after randomly sampling participants with replacement. The resulting distribution was then converted into a z-value at each voxel, which allowed the map to be thresholded based on a corresponding p-value. We used $p < 0.005$ as the threshold to visualize the most reliable weights, which allowed us to assess the face validity of the model (Figure 2A). It is important to note that we did use this thresholded map to perform any inferences. We further computed the spatial intersubject correlation across the models trained on each bootstrapped sample to estimate the approximate reliability of the spatial pattern of weights. This metric can be interpreted similarly to a reliability coefficient, but will be somewhat inflated compared to using completely independent data. Third, we trained the final model using the data from all participants, which is what we ultimately used to test on all other datasets. This model will be the most reliable as it was trained on all available data. For all tests, we used a forced-choice accuracy procedure to evaluate the performance of the model. Forced-choice accuracy examines the relative expressions of the model between the two brain images collected from the same participant and is well suited for fMRI as the input images are unlikely to be on the same scale across individuals or scanners [42,45]. We performed hypothesis tests using permutations in which the labels for each image across participants were randomly flipped 10,000 times to generate a null distribution. We were only interested in whether the target condition was significantly greater than the reference condition, so we reported one-tailed tests. We also computed receiver operator character (ROC) curves using forced choice accuracy. An interesting property of forced choice accuracy is that it is equivalent to sensitivity, specificity, and area under the curve (AUC) of the ROC curve.

Model validation using an independent testing dataset. In order to examine the validity of the trust brain model in an even more rigorous and unbiased way beyond cross-validation, we evaluated its generalizability to a new test dataset (Dataset 2). This dataset was collected in a different country (The Netherlands), using a different scanner and variant of the trust game. We computed forced-choice accuracy on this dataset based on the spatial similarity of the trust model and each participant's trust and distrust beta images estimated using a first level GLM, and also calculated an ROC curve to quantify the tradeoff of sensitivity and specificity at different thresholds (Figure 2B).

Construct validity and specificity of trust model

To evaluate the convergent and discriminant validity of the trust model to other psychological constructs, we tested our trust classification model on other datasets probing distinct psychological constructs, including: beliefs of safety (Dataset 3 and 4), negative affect (Dataset 5 and 6), feelings of anticipated reward (Dataset 7, 8 and 9), cognitive control (Dataset 10), social cognition (Dataset 11 and 12), and self-referential cognition (Dataset 13).

For each dataset, we computed the spatial similarity between the trust multivariate brain pattern and each participant's beta maps representing the test and control conditions from each task. For example, we evaluated how well the trust model could discriminate between safety and risk in datasets 3 and 4, neutral and negative emotional experience in dataset 5 and 6, positive and negative emotional experiences in dataset 6, anticipated reward and loss in dataset 7, anticipated money gain and loss in dataset 8, anticipated reward and no-reward in dataset 9, success and failure in cognitive control in dataset 10, familiarity and unfamiliarity in dataset 11, social and non-social viewing in dataset 12, as well as self and non-self referential cognition in dataset 13. We followed the same forced-choice testing procedure outlined above. Assessing the generalizability of the trust model across different datasets in this manner allowed us to demonstrate convergent and discriminant validity of the trust brain model with other psychological constructs.

Trust Nomological network

Finally, we were interested in assessing the overall spatial similarity between all of the 13 datasets in order to assess the trust nomological network. We employed both qualitative and quantitative approaches. First, to qualitatively visualize the similarity of all of the participants from all 13 datasets (N=547), we used Uniform Manifold Approximation and Projection (UMAP), a nonlinear dimensionality reduction technique (Figure 4A; https://github.com/lmcinnes/umap). UMAP attempts to project high dimensional data into a low dimensional space preserving both local and global distance in the feature space using manifold learning [74]. We first removed dataset-specific differences in brain activity by subtracting the mean brain activity of each dataset from each brain map, ensuring that mean brain activity of each dataset was the same. We used arbitrarily selected values for the hyperparameters (number of neighbors = 50, minimal distance = 0.001). Because the ROI maps contained considerably less features, we used a lower neighbor embedding (Figure S2; number of neighbors = 15, minimal distance = 0.001). Second, to quantitatively assess the overall similarity between the datasets, we averaged beta maps across participants for each condition and computed the spatial similarity across conditions from all datasets in a hierarchical clustered heatmap (Figure 4B).

Acknowledgements

The authors thank Zhihao Zhang and Ifat Levy for sharing their data and Wasita Mahaphanit for her comments on the paper. This research was supported by funding from the Chiang Ching-Kuo Foundation for International Scholarly Exchange (no. GS040-A-16 to P.-H.C.), the Young Scholar Fellowship Program by the Ministry of Science and Technology (MOST 109-2636-H-002-006 and 110-2636-H-002-004 to P.-H.C.) in Taiwan, National Institute of Health grants (nos. R01MH116026 and R56MH080716 to L.J.C., R15MH122927 to D.S.F.), the McKnight Foundation (to M.R.D.), the Netherlands Organization for Scientific Research (NWO; VENI Grant Number 451- 10-021 to B.G.) and the National Science Foundation (NSF CAREER 1848370 to L.J.C.). Code used to perform the analysis in the paper will be available on github and the trust model will be uploaded to Neurovault pending publication of this manuscript.

Supplementary materials

Table S1. Basic information of each dataset as well as forced-choice classification accuracy and p values for each generalization testing.

Dataset number	Number of participants	Construct	Name of the two conditions	Accuracy	P value
1	40	Trust	Trust vs. Distrust	90%	< 0.001
2	17	Trust	Trust vs. Distrust	82%	0.006
3	15	Safety	Safety vs. Risk	93%	< 0.001
4	123	Safety	Safety vs. Risk	93%	< 0.001
5	93	Affect	Neutral vs. Negative	72%	< 0.001
6	56	Affect	Neutral vs. Negative	69%	0.002
6	56	Affect	Positive vs. Negative	73%	< 0.001
6	56	Affect	Positive vs. Neutral	50%	0.551
7	64	Reward	Reward vs. Loss	42%	0.921
8	18	Reward	Gain vs. Loss	66%	0.119
9	29	Reward	Reward vs. No reward	38%	0.933
10	19	Cognitive control	Inhibition success vs. failure	57%	0.326
11	16	Social cognition	Familiar vs. unfamiliar faces	63%	0.217
12	36	Social cognition	Social vs. nonsocial scenes	47%	0.686
13	27	Social cognition	Self vs. nonself referential	40%	0.876

References

1. Arrow, K. *The Limits of Organization*. (citeulike.org, 1974).
2. Dyer, J. H. & Chu, W. The Role of Trustworthiness in Reducing Transaction Costs and Improving Performance: *Organization Science* **14**, 57–68 (2003).
3. Sampson, R. J., Raudenbush, S. W. & Earls, F. Neighborhoods and violent crime: a multilevel study of collective efficacy. *Science* **277**, 918–924 (1997).
4. Knack, S. & Keefer, P. Does Social Capital Have an Economic Payoff? A Cross-Country Investigation. *Q. J. Econ.* **112**, 1251–1288 (1997).
5. Chang, L. The science of trust. (2017).
6. Simpson, J. A. Psychological Foundations of Trust. *Curr. Dir. Psychol. Sci.* **16**, 264–268 (2007).
7. Berg, J., Dickhaut, J. & McCabe, K. Trust, Reciprocity, and Social History. *Games Econ. Behav.* **10**, 122–142 (1995).
8. Dufwenberg, M. & Gneezy, U. Measuring Beliefs in an Experimental Lost Wallet Game. *Games Econ. Behav.* **30**, 163–182 (2000).
9. Camerer, C. *Behavioral Game Theory: Experiments in Strategic Interaction*. (Princeton University Press, 2003).
10. Phan, K. L., Sripada, C. S., Angstadt, M. & McCabe, K. Reputation for reciprocity engages the brain reward center. *Proceedings of the National Academy of Sciences* **107**, 13099–13104 (2010).
11. McCabe, K., Houser, D., Ryan, L., Smith, V. & Trouard, T. A functional imaging study of cooperation in two-person reciprocal exchange. *Proc. Natl. Acad. Sci. U. S. A.* **98**, 11832–11835 (2001).
12. Fareri, D. S., Chang, L. J. & Delgado, M. R. Computational Substrates of Social Value in Interpersonal Collaboration. *Journal of Neuroscience* **35**, 8170–8180 (2015).
13. Delgado, M. R., Frank, R. H. & Phelps, E. A. Perceptions of moral character modulate the neural systems of reward during the trust game. *Nat. Neurosci.* **8**, 1611–1618 (2005).
14. Fareri, D. S., Chang, L. J. & Delgado, M. R. Effects of direct social experience on trust decisions and neural reward circuitry. *Front. Neurosci.* **6**, (2012).
15. Krajbich, I., Adolphs, R., Tranel, D., Denburg, N. L. & Camerer, C. F. Economic games quantify diminished sense of guilt in patients with damage to the prefrontal cortex. *J. Neurosci.* **29**, 2188–2192 (2009).
16. King-Casas, B. *et al.* Getting to know you: reputation and trust in a two-person economic exchange. *Science* **308**, 78–83 (2005).
17. King-Casas, B. *et al.* The rupture and repair of cooperation in borderline personality disorder. *Science* **321**, 806–810 (2008).
18. Fouragnan, E. *et al.* Reputational priors magnify striatal responses to violations of trust. *J. Neurosci.* **33**, 3602–3611 (2013).
19. Chang, L. J., Smith, A., Dufwenberg, M. & Sanfey, A. G. Triangulating the neural, psychological, and economic bases of guilt aversion. *Neuron* **70**, 560–572 (2011).
20. van Baar, J. M., Chang, L. J. & Sanfey, A. G. The computational and neural substrates of moral strategies in social decision-making. *Nat. Commun.* **10**, 1483 (2019).
21. Frank, R. H., Gilovich, T. & Regan, D. T. The evolution of one-shot cooperation: An experiment. *Ethol. Sociobiol.* **14**, 247–256 (1993).
22. Jolly, E. & Chang, L. J. Gossip drives vicarious learning and facilitates robust social connections. (2018) doi:10.31234/osf.io/qau5s.
23. Feinberg, M., Willer, R. & Schultz, M. Gossip and ostracism promote cooperation in groups. *Psychol. Sci.* **25**, 656–664 (2014).
24. Sommerfeld, R. D., Krambeck, H.-J., Semmann, D. & Milinski, M. Gossip as an alternative

for direct observation in games of indirect reciprocity. *Proc. Natl. Acad. Sci. U. S. A.* **104**, 17435–17440 (2007).

25. Stanley, D. A., Sokol-Hessner, P., Banaji, M. R. & Phelps, E. A. Implicit race attitudes predict trustworthiness judgments and economic trust decisions. *Proc. Natl. Acad. Sci. U. S. A.* **108**, 7710–7715 (2011).
26. Scharlemann, J. P. W., Eckel, C. C., Kacelnik, A. & Wilson, R. K. The value of a smile: Game theory with a human face. *J. Econ. Psychol.* **22**, 617–640 (2001).
27. van 't Wout, M. & Sanfey, A. G. Friend or foe: the effect of implicit trustworthiness judgments in social decision-making. *Cognition* **108**, 796–803 (2008).
28. Scanzoni, J. Social exchange and behavioral interdependence. *Social exchange in developing relationships* 61–98 (1979).
29. Rousseau, D. M., Sitkin, S. B., Burt, R. S. & Camerer, C. Not So Different After All: A Cross-Discipline View Of Trust. *Acad. Manage. Rev.* **23**, 393–404 (1998).
30. Chang, L. J., Doll, B. B., van't Wout, M., Frank, M. J. & Sanfey, A. G. Seeing is believing: Trustworthiness as a dynamic belief. *Cogn. Psychol.* **61**, 87–105 (2010).
31. Rilling, J. *et al.* A neural basis for social cooperation. *Neuron* **35**, 395–405 (2002).
32. Kishida, K. T. & Montague, P. R. Imaging models of valuation during social interaction in humans. *Biol. Psychiatry* **72**, 93–100 (2012).
33. Tabibnia, G., Satpute, A. B. & Lieberman, M. D. The Sunny Side of Fairness: Preference for Fairness Activates Reward Circuitry (and Disregarding Unfairness Activates Self-Control Circuitry). *Psychol. Sci.* **19**, 339–347 (2008).
34. Bohnet, I. & Zeckhauser, R. Trust, risk and betrayal. *J. Econ. Behav. Organ.* **55**, 467–484 (2004).
35. Cronbach, L. J. & Meehl, P. E. Construct validity in psychological tests. *Psychol. Bull.* **52**, 281–302 (1955).
36. Campbell, D. T. & Fiske, D. W. Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychol. Bull.* **56**, 81–105 (1959).
37. Woo, C.-W., Chang, L. J., Lindquist, M. A. & Wager, T. D. Building better biomarkers: brain models in translational neuroimaging. *Nat. Neurosci.* **20**, 365–377 (2017).
38. Poldrack, R. A., Halchenko, Y. O. & Hanson, S. J. Decoding the large-scale structure of brain function by classifying mental States across individuals. *Psychol. Sci.* **20**, 1364–1372 (2009).
39. Yarkoni, T., Poldrack, R. A., Nichols, T. E., Van Essen, D. C. & Wager, T. D. Large-scale automated synthesis of human functional neuroimaging data. *Nat. Methods* **8**, 665–670 (2011).
40. Chang, L. J., Yarkoni, T., Khaw, M. W. & Sanfey, A. G. Decoding the role of the insula in human cognition: functional parcellation and large-scale reverse inference. *Cereb. Cortex* **23**, 739–749 (2013).
41. Kragel, P. A. *et al.* Generalizable representations of pain, cognitive control, and negative emotion in medial frontal cortex. *Nat. Neurosci.* **21**, 283–289 (2018).
42. Wager, T. D. *et al.* An fMRI-based neurologic signature of physical pain. *N. Engl. J. Med.* **368**, 1388–1397 (2013).
43. Krishnan, A. *et al.* Somatic and vicarious pain are represented by dissociable multivariate brain patterns. *Elife* **5**, (2016).
44. Eisenbarth, H., Chang, L. J. & Wager, T. D. Multivariate Brain Prediction of Heart Rate and Skin Conductance Responses to Social Threat. *J. Neurosci.* **36**, 11987–11998 (2016).
45. Chang, L. J., Gianaros, P. J., Manuck, S. B., Krishnan, A. & Wager, T. D. A Sensitive and Specific Neural Signature for Picture-Induced Negative Affect. *PLoS Biol.* **13**, e1002180 (2015).
46. Yu, H. *et al.* A Generalizable Multivariate Brain Pattern for Interpersonal Guilt. *Cereb. Cortex* **30**, 3558–3572 (2020).

47. Nishimoto, S. *et al.* Reconstructing visual experiences from brain activity evoked by natural movies. *Curr. Biol.* **21**, 1641–1646 (2011).
48. Huth, A. G. *et al.* Decoding the Semantic Content of Natural Movies from Human Brain Activity. *Front. Syst. Neurosci.* **10**, 81 (2016).
49. Gorgolewski, K., Esteban, O., Schaefer, G., Wandell, B. & Poldrack, R. OpenNeuro—a free online platform for sharing and analysis of neuroimaging data. *Organization for human brain mapping. Vancouver, Canada* **1677**, (2017).
50. Gorgolewski, K. J. *et al.* NeuroVault.org: a web-based repository for collecting and sharing unthresholded statistical maps of the human brain. *Front. Neuroinform.* **9**, 8 (2015).
51. Yarkoni, T., Poldrack, R. A., Van Essen, D. C. & Wager, T. D. Cognitive neuroscience 2.0: building a cumulative science of human brain function. *Trends Cogn. Sci.* **14**, 489–496 (2010).
52. Poldrack, R. A. & Gorgolewski, K. J. Making big data open: data sharing in neuroimaging. *Nat. Neurosci.* **17**, 1510–1517 (2014).
53. Gorgolewski, K. J. *et al.* The brain imaging data structure, a format for organizing and describing outputs of neuroimaging experiments. *Sci Data* **3**, 160044 (2016).
54. Woo, C.-W. *et al.* Separate neural representations for physical pain and social rejection. *Nat. Commun.* **5**, 5380 (2014).
55. Schreuders, E., Klapwijk, E. T., Will, G.-J. & Güroğlu, B. Friend versus foe: Neural correlates of prosocial decisions for liked and disliked peers. *Cogn. Affect. Behav. Neurosci.* **18**, 127–142 (2018).
56. Schonberg, T. *et al.* Decreasing ventromedial prefrontal cortex activity during sequential risk-taking: an fMRI investigation of the balloon analog risk task. *Front. Neurosci.* **6**, (2012).
57. Poldrack, R. A. *et al.* A phenome-wide examination of neural and cognitive function. *Sci Data* **3**, 160110 (2016).
58. Chen, P.-H. A. & Wu, Y.-J. C. Uncovering the association between individual variations in self-awareness and brain emotional reactivity by intersubject representational analysis. *in preparation*.
59. Barch, D. M. *et al.* Function in the human connectome: Task-fMRI and individual differences in behavior. *Neuroimage* **80**, 169–189 (2013).
60. Zhang, Z. *et al.* Distributed neural representation of saliency controlled value and category during anticipation of rewards and punishments. *Nat. Commun.* **8**, 447 xii (2017).
61. Tomova, L. *et al.* Acute social isolation evokes midbrain craving responses similar to hunger. *Nat. Neurosci.* (2020) doi:10.1101/2020.03.25.006643.
62. Xue, G., Aron, A. R. & Poldrack, R. A. Common neural substrates for inhibition of spoken and manual responses. *Cereb. Cortex* **18**, 1923–1932 (2008).
63. Wakeman, D. G. & Henson, R. N. A multi-subject, multi-modal human neuroimaging dataset. *Sci Data* **2**, 150001 (2015).
64. Chen, P.-H. A., Kelley, W. M., Lopez, R. B. & Heatherton, T. F. Reducing reward responsivity and daily food desires in female dieters through domain-specific training. *Soc. Neurosci.* **14**, 470–483 (2019).
65. Chen, P.-H. A., Wagner, D. D., Kelley, W. M. & Heatherton, T. F. Activity in cortical midline structures is modulated by self-construal changes during acculturation. *Culture and Brain* **3**, 39–52 (2015).
66. Chen, P.-H. A., Wagner, D. D., Kelley, W. M., Powers, K. E. & Heatherton, T. F. Medial prefrontal cortex differentiates self from mother in Chinese: evidence from self-motivated immigrants. *Culture and Brain* **1**, 3–15 (2013).
67. Jolly, E. & Chang, L. J. Multivariate Spatial Feature Selection in fMRI. *Soc. Cogn. Affect. Neurosci.* (2021) doi:10.1093/scan/nsab010.
68. Krueger, F. *et al.* Neural correlates of trust. *Proceedings of the National Academy of Sciences* **104**, 20084–20089 (2007).

69. Geier, C. F., Terwilliger, R., Teslovich, T., Velanova, K. & Luna, B. Immaturities in reward processing and its influence on inhibitory control in adolescence. *Cereb. Cortex* **20**, 1613–1629 (2010).
70. Delgado, M. R., Nystrom, L. E., Fissell, C., Noll, D. C. & Fiez, J. A. Tracking the hemodynamic responses to reward and punishment in the striatum. *J. Neurophysiol.* **84**, 3072–3077 (2000).
71. Knutson, B., Westdorp, A., Kaiser, E. & Hommer, D. FMRI visualization of brain activity during a monetary incentive delay task. *Neuroimage* **12**, 20–27 (2000).
72. Venkatraman, V., Payne, J. W., Bettman, J. R., Luce, M. F. & Huettel, S. A. Separate neural mechanisms underlie choices and strategic preferences in risky decision making. *Neuron* **62**, 593–602 (2009).
73. Rand, D. G., Greene, J. D. & Nowak, M. A. Spontaneous giving and calculated greed. *Nature* **489**, 427–430 (2012).
74. McInnes, L., Healy, J. & Melville, J. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. *arXiv [stat.ML]* (2018).
75. Kriegeskorte, N. & Douglas, P. K. Interpreting encoding and decoding models. *Curr. Opin. Neurobiol.* **55**, 167–179 (2019).
76. Sul, S. *et al.* Spatial gradient in value representation along the medial prefrontal cortex reflects individual differences in prosociality. *Proc. Natl. Acad. Sci. U. S. A.* **112**, 7851–7856 (2015).
77. Haufe, S. *et al.* On the interpretation of weight vectors of linear models in multivariate neuroimaging. *Neuroimage* **87**, 96–110 (2014).
78. Jolly, E. & Chang, L. J. Multivariate spatial feature selection in fMRI. *osf.io*.
79. Chen, P.-H. A., Jolly, E., Cheong, J. H. & Chang, L. J. Intersubject representational similarity analysis reveals individual variations in affective experience when watching erotic movies. *Neuroimage* **216**, 116851 (2020).
80. Finn, E. S. *et al.* Idiosynchrony: From shared responses to individual differences during naturalistic neuroimaging. *Neuroimage* **215**, 116828 (2020).
81. Jolly, E. & Chang, L. J. The Flatland Fallacy: Moving Beyond Low–Dimensional Thinking. *Top. Cogn. Sci.* (2019).
82. Poldrack, R. A. Can cognitive processes be inferred from neuroimaging data? *Trends Cogn. Sci.* **10**, 59–63 (2006).
83. Misaki, M., Kim, Y., Bandettini, P. A. & Kriegeskorte, N. Comparison of multivariate classifiers and response normalizations for pattern-information fMRI. *Neuroimage* **53**, 103–118 (2010).
84. Gorgolewski, K. J., Durnez, J. & Poldrack, R. A. Preprocessed Consortium for Neuropsychiatric Phenomics dataset. *F1000Res.* **6**, 1262 (2017).
85. Gianaros, P. J. *et al.* An inflammatory pathway links atherosclerotic cardiovascular disease risk to neural activity evoked by the cognitive regulation of emotion. *Biol. Psychiatry* **75**, 738–745 (2014).
86. Wagner, D. D. & Heatherton, T. F. Self-regulatory depletion increases emotional reactivity in the amygdala. *Soc. Cogn. Affect. Neurosci.* **8**, 410–417 (2012).
87. Glasser, M. F. *et al.* The minimal preprocessing pipelines for the Human Connectome Project. *Neuroimage* **80**, 105–124 (2013).
88. Krebs, R. M., Heipertz, D., Schuetze, H. & Duzel, E. Novelty increases the mesolimbic functional connectivity of the substantia nigra/ventral tegmental area (SN/VTA) during reward anticipation: Evidence from high-resolution fMRI. *Neuroimage* **58**, 647–655 (2011).
89. Esteban, O. *et al.* fMRIPrep: a robust preprocessing pipeline for functional MRI. *Nat. Methods* **16**, 111–116 (2019).
90. Powers, K. E., Wagner, D. D., Norris, C. J. & Heatherton, T. F. Socially excluded individuals fail to recruit medial prefrontal cortex for negative social scenes. *Soc. Cogn. Affect.*

Neurosci. (2011).

91. Wagner, D. D., Boswell, R. G., Kelley, W. M. & Heatherton, T. F. Inducing Negative Affect Increases the Reward Value of Appetizing Foods in Dieters. *J. Cogn. Neurosci.* **24**, 1625–1633 (2012).
92. Wagner, D. D., Altman, M., Boswell, R. G., Kelley, W. M. & Heatherton, T. F. Self-Regulatory Depletion Enhances Neural Responses to Rewards and Impairs Top-Down Control. *Psychol. Sci.* **24**, 2262–2271 (2013).