PNAS RESEARCH ARTICLE | NEUROSCIENCE
PSYCHOLOGICAL AND COGNITIVE SCIENCES

Distinguishing deception from its confounds by improving the validity of fMRI-based neural prediction

Sangil Lee[a,1], Runxuan Niu[b], Lusha Zhu[b,1], Andrew S. Kayser[c,d,1], and Ming Hsu[a,e,1]

Affiliations are included on p. 10.

Edited by Read Montague, Fralin Biomedical Research Institute, Virginia Tech Carilion Research Institute, Roanoke, VA; received June 26, 2024; accepted October 22, 2024 by Editorial Board Member Michael S. Gazzaniga

Deception is a universal human behavior. Yet longstanding skepticism about the validity of measures used to characterize the biological mechanisms underlying deceptive behavior has relegated such studies to the scientific periphery. Here, we address these fundamental questions by applying machine learning methods and functional magnetic resonance imaging (fMRI) to signaling games capturing motivated deception in human participants. First, we develop an approach to test for the presence of confounding processes and validate past skepticism by showing that much of the predictive power of neural predictors trained on deception data comes from processes other than deception. Specifically, we demonstrate that discriminant validity is compromised by the predictor's ability to predict behavior in a control task that does not involve deception. Second, we show that the presence of confounding signals need not be fatal and that the validity of the neural predictor can be improved by removing confounding signals while retaining those associated with the task of interest. To this end, we develop a "dual-goal tuning" approach in which, beyond the typical goal of predicting the behavior of interest, the predictor also incorporates a second compulsory goal that enforces chance performance in the control task. Together, these findings provide a firmer scientific foundation for understanding the neural basis of a neglected class of behavior, and they suggest an approach for improving validity of neural predictors.

deception | lie detection | MVPA | prediction | validity

Significance

Complex behaviors such as deception engage a myriad of co-occurring processes, making it difficult to determine whether any predictor of deception actually predicts deceptive processes, per se. Here, we confirm this problem by demonstrating that a neural predictor of deception can also predict selfish, nondeceptive choices, suggesting that the predictor utilizes other signals. To remediate this problem, we develop a machine-learning method that pursues "dual goals" to ensure that the predictor identifies behaviors of interest but not confounding behaviors from a parallel experimental task. In addition to its application to deception, we argue that this approach may permit improved isolation of other complex cognitive processes and thereby open the door to more sophisticated questions addressing the dissociability of cognitive constructs.

Claims that scientifically based methods can detect deception have been made since at least the early part of the 20th century. Yet from the start, such proposals, typically driven by forensic goals, have been met with intense skepticism from the scientific community, in large part due to uncertainty about the nature of the measured signals (1–4). Indeed, a conundrum dating to the earliest attempts at detecting deception concerns how to rule out the myriad alternative processes, such as those involved in arousal, weighing of risks and rewards, and belief inference, that often co-occur with deception but are not necessarily themselves indicative of deception (3–6).

Scientifically, this continued lack of progress in our ability to assess and exclude validity threats has severely impeded progress in understanding the neural bases of deceptive and honest behaviors, which are central in studies of mate selection, social communication, and economic exchange, among others (7–10). As described by the National Research Council, "An indication of the state of the field is the fact that the validity questions that scientists raise today include many of the same ones that were first articulated in criticisms of Marston's original work in 1917" (4).

In recent years, however, there has been renewed excitement about the possibility of a more scientifically grounded understanding of the neural basis of deception. New analysis approaches offer formal, testable ways to decode mental states from brain data (11, 12) due to a confluence of advances in behavioral, neural, and statistical methods, and, in particular, the application of machine learning pattern analysis techniques to economic signaling games (13–16). Such games have now been extensively studied in the economic and biological sciences in order to model goal-directed communication between agents, including the possibility for motivated deception (7–9).

Here, we seek to build upon these advances by developing a set of methodological and statistical tools to enable researchers to systematically test and improve the validity of putative predictors of deception. As a first step toward establishing criterion validity, we tested the accuracy of a neural predictor by applying multivariate decoding methods to functional neuroimaging data while participants made decisions about whether to send honest or deceptive messages to another participant. Our initial results show that once trained on behavior in this game, a whole-brain neural predictor is capable of distinguishing

Author contributions: L.Z., A.S.K., and M.H. designed research; R.N. and L.Z. performed research; S.L. contributed new reagents/analytic tools; S.L., R.N., A.S.K., and M.H. analyzed data; and S.L., R.N., L.Z., A.S.K., and M.H. wrote the paper.

The authors declare no competing interest.

This article is a PNAS Direct Submission. R.M. is a guest editor invited by the Editorial Board.

[1]To whom correspondence may be addressed. Email: sangillee3rd@gmail.com, lushazhu@pku.edu.cn, andrew.kayser@ucsf.edu, or mhsu@haas.berkeley.edu.

This article contains supporting information online at https://www.pnas.org/lookup/suppl/doi:10.1073/pnas.2412881121/-/DCSupplemental.

Published December 6, 2024.

between deceptive and honest behavior at rates significantly higher than chance.

Second, and more importantly, we set out to address questions about the construct validity of our neural predictor—i.e., how and why our predictor works. Whereas accuracy metrics ask whether our neural predictor can make accurate predictions, construct validity asks whether our neural predictor is truly measuring the underlying process it purports to measure. Despite its acknowledged importance, and the many methods of detection proposed over the past 100 y, there has been surprisingly little effort in attempting to assess the construct validity of these methods, nor even an agreement on what constitutes scientific evidence for or against the presence of validity threats (3, 4, 17).

In particular, we focus on the aspect of construct validity that pertains to discriminant validity (18), which assesses the extent to which our deception predictor is driven by processes not specific to deception. For this reason, we introduced a second, isomorphic game in which players can achieve the same ends (i.e., payoffs) as in the deception game, but via nondeceptive means. That is, the two games share the same players, strategies, and payoffs such that only surface labels (i.e., the messages) differ (19). Critically, if the putative deception predictor "overgeneralizes"—meaning that its predictions are also correlated with signals underlying the control game—this result would provide strong evidence that its predictive power is significantly driven by processes held in common between the two games—e.g., self-interested motives, belief inference, arousal associated with violation of social norms, or others—rather than those specific to deception itself (4, 6, 12, 20).

Using the neural predictor trained on the deception game, we show that its construct validity is significantly compromised by the fact that this predictor also predicts "merely selfish" (i.e., selfish but not deceptive) behavior in the control game. Indeed, the magnitude of this effect is such that the prediction rate in the control task is statistically indistinguishable from that for the task of interest. Moreover, performance falls to chance when the predictor is asked to distinguish between a) deceptive choices and b) selfish but not deceptive choices. Finally, at the neural level, we find that overgeneralization is pervasive across the brain such that many regions that predict deception are at least partially driven by signals common to the control task.

Having identified the presence of confounding processes, we further investigated the extent to which the influence of any identified confounding processes can be removed or mitigated. This step is particularly important because the presence of confounding signals need not rule out the possible presence of a coexisting deception-specific signal. If so, we may be able to improve the validity of the neural predictor by purging the set of signals common to both tasks. To this end, we develop a statistical approach where, in addition to the typical goal of predicting the behavior of interest, the predictor also incorporates a second goal enforcing chance performance in the control task.

We show that compared to three other potential alternative methods considered, this "dual-goal tuning" approach is able to construct a whole-brain deception predictor that predicts deception but does not rely on neural patterns held in common with the control task, rendering it capable of distinguishing between deceptive and merely selfish behavior. Additionally, this method uncovers substantial variation in the extent to which deception-specific signals can be recovered: Whereas dual-goal tuning in some regions, such as the primary visual cortex, entirely purges presumptive predictive signals—suggesting that predictive accuracy in these regions is driven by processes other than deception—other regions, including ones previously implicated in meta-analyses of deception (such as the superior anterior cingulate cortex and superior frontal gyrus), retain significant predictive power after correction. Together, these findings potentially enhance the scientific rigor of studies that assess the neural basis of deceptive and honest behaviors, and they represent an important step forward in building a firmer scientific foundation necessary for ongoing progress in detecting deception in forensic settings (1–5).

Results

Signaling Game Approach to Dissociate Processes Underlying Deception. To identify a set of processes that co-occur with, but are not necessarily indicative of, deception, we used a signaling game approach that has been extensively employed in behavioral economics and evolutionary biology to capture the role of communication in economic interactions, including tradeoffs between honesty and deception (7–10, 19). Specifically, we designed a pair of isomorphic signaling games that differed only in the extent to which players' actions could be assigned a truth value (8–10). In both games, the participant (the sender) is presented with two potential allocations of monetary gains for themselves and a counterpart (the receiver). Importantly, participants are informed that as the sender, they can see the options but cannot make the choice between them, while the receiver cannot see the options but is responsible for the choice. This manipulation thus renders the receiver completely reliant upon the sender for any information about the choice. On each trial, one allocation provides a larger payoff to the sender, while the other provides a larger payoff to the receiver. Critically, in the deception task, senders must choose between two messages that are verifiably truthful or false (e.g., "Option A will earn you more money than B"; Fig. 1*A*—deception task). In contrast, in the control task, the senders' messages do not have a truth value (e.g., "I prefer that you choose option A"; Fig. 1*A*—control task). Participants were not given any feedback about whether their chosen option was accepted by the receiver, but they were told that based on previous studies, receivers tend to choose the suggested option 78% of the time. Additionally, participants' bonus payments were determined by randomly picking one trial of each task and carrying out the selected message with 78% probability (see *SI Appendix* for further task details).

This design incorporates two important features that together help to identify deception-specific processes. First, in contrast with previous paradigms involving instructed lies (20–24), behavior in signaling games captures the idea that honesty and deception are properties of the communicative signals that agents send to one another in the service of some economic or evolutionary value. Second, and more importantly, the inclusion of an isomorphic control game allows us to identify the set of processes that are also present in other nondeceptive decisions—for example, those associated with weighing costs and benefits to oneself, or concerns for fairness—and thus that are not specific to deception per se.

Consistent with previous findings showing that processes underlying deception can be dissociated from those underlying other types of norm violations (8–10), there was a significant difference in how participants behaved in the two conditions. In particular, the need to send a deceptive message reduced the proportion of messages recommending the selfish option to the receiver (54.6%) compared to the control condition (61.9%; paired t test $P = 0.0074$). Notably, individual differences in sensitivity to social norms around self-favoring and honest responses could be dissociated across the two tasks, indicating that the message manipulation differentially impacted behavior: Those participants who made more deceptive choices in the deception task were not necessarily the ones who made more selfish choices in the control task (Fig. 1*B*; $r = 0.28$, $P = 0.11$).

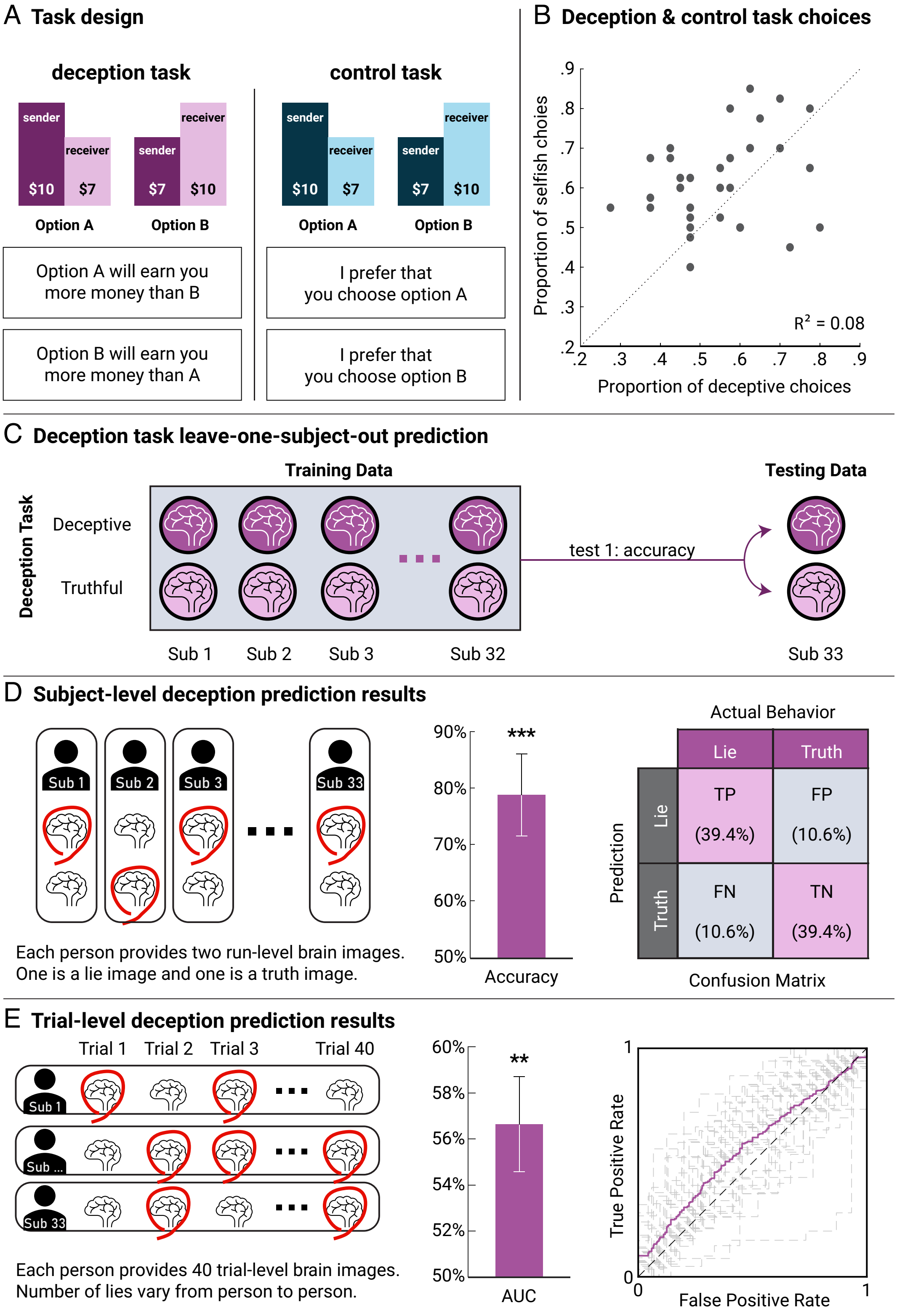


Fig. 1. (*A*) Experimental task design dissociating deception from basic social decision-making processes. Both task conditions (deception and control) involve allocations of monetary gains between the participant and a counterpart. Participants always played the role of the message sender, who sends one of two messages to the receiver. The deception task requires senders to choose between two messages that are verifiably truthful or false (e.g., “Option A will earn you more money than B”). In the Control Task, the senders’ messages do not have a truth value (e.g., “I prefer that you choose option A”). (*B*) Scatterplot of the proportion of deceptive choices from the main task and the proportion of selfish choices in the control task. That behavior in the deception task could be partially dissociated from that in the control task suggests a contribution of processes that are not simply reducible to those captured in the control task. (*C*) Schematic of test 1: leave-one-out cross-validation procedure for within-task prediction. (*D*) Predictors were trained on subjects’ run-level average neural activity for each choice type (two images per task for each subject). Neural predictors of deception predicted deceptive behavior at rates significantly greater than chance (78.8% ± 7.24% (mean ± SE), $P < 0.001$). (*E*) In trial-level prediction, each trial’s activity was estimated separately and predictions were made at the trial level. The neural predictor of trial-level deception also showed significant prediction of deceptive choices (AUC = 56.6% ± 2.1% (mean ± SE), $P = 0.004$). **$P < 0.01$ and ***$P < 0.001$.

Neural Predictor Distinguishing between Deceptive and Honest Behavior. Next, using functional neuroimaging data, we sought to assess the extent to which a whole-brain neural predictor, once trained on neural responses associated with deceptive and honest behavior (Fig. 1*C*), could predict deception in holdout data using brain activity alone (13). Specifically, training and testing were performed at both subject and trial levels. In subject-level prediction, we estimated, for each subject, one image associated with truthful behavior and one with deceptive behavior, by estimating average brain activity across the same trial types (25, 26). We found that the neural predictor was able to correctly distinguish the two images at rates significantly greater than chance (78.8%, P < 0.001; Fig. 1*D*). In trial-level prediction, separate images were estimated for each choice (27–29). Because the numbers of deceptive and honest choices are not balanced, a simple model that predicts the most frequent choice can achieve accuracy higher than 50%. We therefore used the area under the receiver operating characteristic curve (AUC) as a measure of overall performance of the classifier, where 50% corresponds to chance performance and 100% corresponds to perfect classification. The AUC can be understood as the accuracy obtained by repeatedly randomly selecting two trials and assessing whether the predictor assigns higher scores to the correct category. We found that the neural predictor performed significantly better than chance at the trial level (average AUC = 56.6%, P = 0.004; Fig. 1*E*).

Significant Presence of Confounding Signals. Although this neural predictor of deception was able to show significant discrimination between deceptive and honest behavior, it is possible that at least some of the predictive signal is not related to deception, but rather to confounding processes. We sought to test for this possibility by building upon recent methods that ask whether a neural predictor of a particular cognitive state shares signals with other processes (12, 29, 30). For example, researchers might develop a neural predictor in one dataset and then apply that predictor to another dataset that utilizes a different task in order to (putatively) assess the same underlying construct. A significant correlation between the predictive signal and the behavior of the new task (i.e., generalization) argues that the predictor of one construct incorporates the other construct, whereas a lack of generalization suggests independence of the constructs in question.

By testing the extent to which the neural predictor of deception trained on the deception task could also distinguish between behavior in the isomorphic control signaling game that did not involve deception, this approach provides one way to identify threats to discriminant validity in the neural predictor (Fig. 2*A*)—in other words, by evaluating whether measures that should not be correlated are indeed uncorrelated (31). In contrast, if a neural predictor trained on the deception game produces signals that are also associated with nondeceptive behaviors, we can conclude that this "naïve" predictor is at least partially driven by processes held in common between the two games—e.g., self-interested motives, belief inference, arousal associated with violation of social norms, or others—rather than those specific to deception itself.

We found strong evidence that the deception predictor incorporates signals that are shared with the control task. Specifically, the neural signal produced by applying a deception predictor to a control task was positively correlated with selfish but nondeceptive choices in the control task (subject-level: r = 0.39, P = 0.021, Fig. 2*B*; trial-level r = 0.084, P = 0.014, Fig. 2*C*). Thus, despite the significant predictive accuracy in the deception task, this overgeneralization provides strong evidence of a lack of discriminant validity for this predictor (Fig. 2*D*). Indeed, the strength of this overgeneralization is such that if one were to use the putative deception predictor to classify the choices in the control task, the prediction accuracy would be statistically indistinguishable from its performance in the deception task at both the subject prediction level (78.8% vs. 69.7%; P = 0.45) and the single-trial prediction level (56.6% vs. 55.4%; P = 0.71).

Methods to Control for Confounding Processes. While the presence of overgeneralization shows the vulnerability of our neural predictor to confounding, it also offers an opportunity to improve its validity by identifying the set of such signals that can be removed if handled appropriately. To explore this possibility, we considered four potential solutions by incorporating data from the control task into the training process (Fig. 2*E*).

First, if the loci containing confounded signals are spatially separate from those carrying signals of interest, a "region removal" method can be used to improve the validity of the neural predictor by removing regions that carry strong confounding signals. Second, we considered a data "relabeling" method, also known as the binary relevance method (32), that is often used in the machine-learning literature in the context of multiclass prediction problems. It has also been used in previous MVPA studies to show dissociability of related cognitive constructs (30). In our setting, this method works by assigning all trials of the control task to be "truth" trials, with the only "lie" trials supplied by the main task. By explicitly defining both control categories as "truth" trials, this method ensures that the predictor is trained to down-weight any confounding signals present in the control task categories (i.e., selfish vs. altruistic). Third, we considered a "regress-out" method, which attempts to focus the neural predictor on signal variation unique to the deception task. This method takes advantage of the isomorphic nature of our task, in that both tasks have the same number of trials, and choices in the two tasks can be paired. As such, we can regress behavioral variation in the control task out of the deception task (see *SI Appendix* for more details on various regress-out methods).

Finally, we developed a fourth method that seeks to directly control for cross-task (over-) generalization, akin to recent efforts in machine learning to incorporate a guiding cost function to better identify "correct" signals (33, 34). However, because the correct signal in our case—i.e., the neural signature of deception—is unknown in its form or loci, we are not able to augment the cost function with additional terms that guide the validity of the signal, as in multiobjective optimization (35). Instead, we approach this problem as a constrained optimization problem by incorporating a negative guiding cost function that penalizes the presence of "incorrect" signals (36).

Formally, let there be two datasets, ($\boldsymbol{X}_1$, $\boldsymbol{Y}_1$) and ($\boldsymbol{X}_2$, $\boldsymbol{Y}_2$), where $\boldsymbol{X}$ is an $n \times p$ matrix of brain activity (n observations, p voxels), and $\boldsymbol{Y}$ an $n \times 1$ column vector of participants' choices. Shared signal is present if the vector of weights $\boldsymbol{b}$ in the linear predictor $\boldsymbol{X}_1\boldsymbol{b}$ of $\boldsymbol{Y}_1$ can also be used to construct a linear predictor $\boldsymbol{X}_2\boldsymbol{b}$ that is predictive of $\boldsymbol{Y}_2$. If both $\boldsymbol{X}$ and $\boldsymbol{Y}$ have been mean-centered, this test can be expressed as follows, in the case of positive over-generalization:

$$cov\left(\boldsymbol{X}_2\boldsymbol{b}, \boldsymbol{Y}_2\right) = \frac{1}{n}\left(\boldsymbol{b}^T \boldsymbol{X}_2^T \boldsymbol{Y}_2\right) = \frac{\boldsymbol{b}^T \boldsymbol{C}_2}{n} > 0, \qquad [1]$$

where $\boldsymbol{C}_2$ is the brain–behavior covariance in Dataset S2. Thus, one way to control cross-task generalization is to introduce $\boldsymbol{X}_2$and $\boldsymbol{Y}_2$ into model training and incorporate [1] into the cost function. Extending the ridge regression cost function, for example, results in:

$$\ln\left(\frac{p\left(\boldsymbol{Y}_1 = 1\right)}{p(\boldsymbol{Y}_1 = 0)}\right) = b_0 + \boldsymbol{X}_1\boldsymbol{b} + \lambda\boldsymbol{b}^T\boldsymbol{b} + \omega\boldsymbol{b}^T\boldsymbol{C}_2, \qquad [2]$$

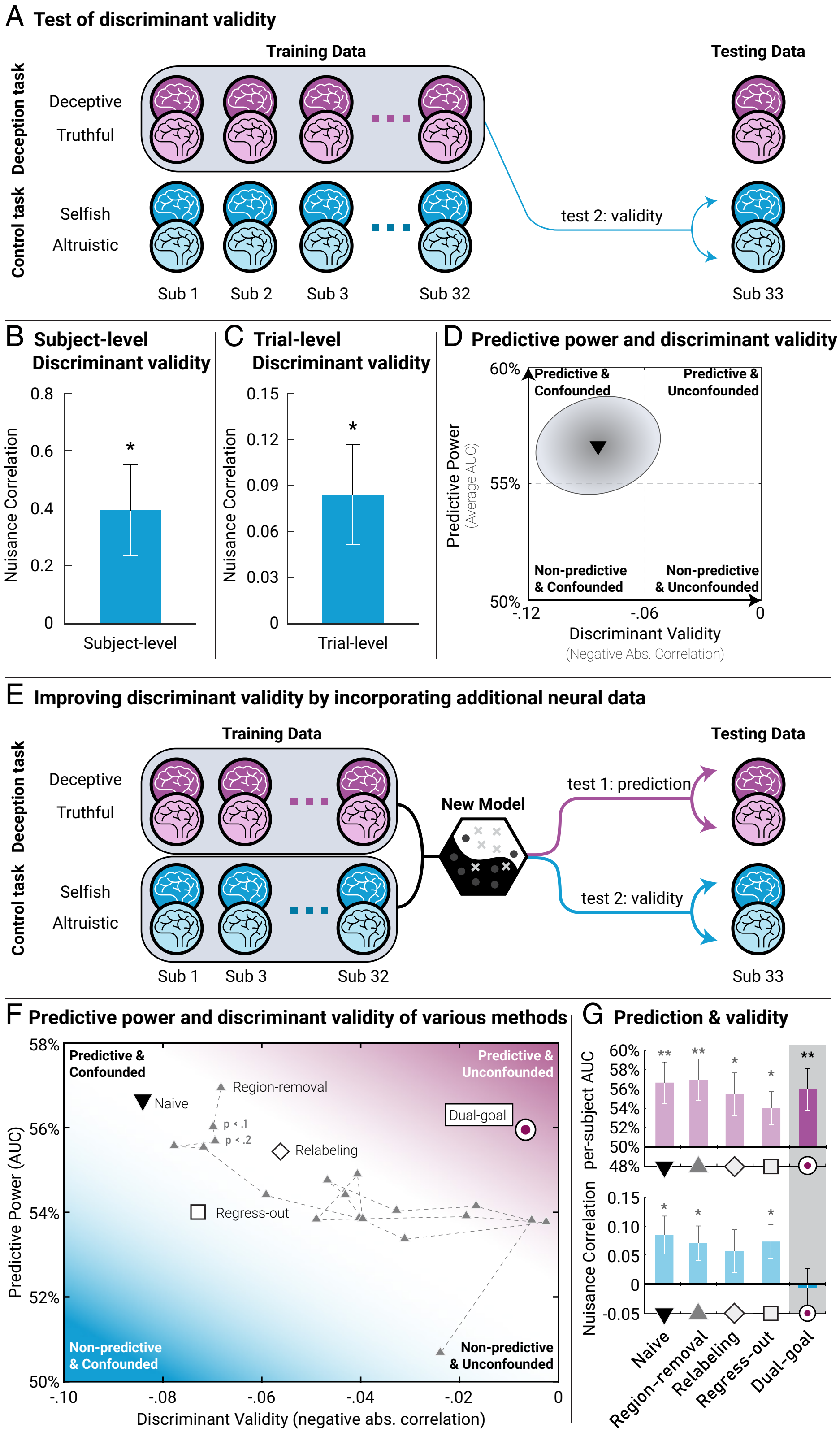


Fig. 2. (*A*) Schematic of test 2. Discriminant validity is tested by using the control task as testing data, rather than deception task trials as in test 1. A predictor that also predicts behavior in the control task indicates a lack of discriminant validity. (*B*) In subject-level testing, the neural predictor of deception showed a lack of discriminant validity, as it was significantly correlated with behavior in the control game (mean r = 0.39, *P* = 0.021). (*C*) Neural predictors showed a lack of discriminant validity in trial-level prediction as well (mean r = 0.084, *P* = 0.014). (*D*) Combining test 1 and test 2, an ideal neural predictor should be located in the *Upper Right* quadrant where it is predictive of deception (ordinate) but uncorrelated with other behaviors (abscissa). In contrast, the neural predictor trained on deception data was located in the *Upper Left* quadrant—i.e., it was able to demonstrate predictive accuracy when tested on deceptive and honest trials, but underlying signals contained processes held in common with the control task. (*E*) Constructing and comparing methods that seek to improve discriminant validity by incorporating the control task into training data. (*F*) As in (*D*), predictive accuracy vs. discriminant validity, compared across different neural predictors. Generally, all methods used, when compared with the naive predictor, showed some tradeoff between predictive power and discriminant validity, with dual-goal tuning providing the most favorable tradeoff in terms of power loss and discriminant validity. (*G*) Bar graphs of prediction performances and validity tests for methods shown in panel (*F*). **P* < 0.05, ***P* < 0.01, and ****P* < 0.001.

which in addition to the standard ridge penalty λ includes a new hyperparameter ω to control for cross-task generalization. That is, the set of hyperparameters (λ, ω) should be chosen to maximize prediction performance while cross-task generalization is held at null. However, the above formulation can be computationally taxing due to the increased parameter space, and difficult to use with approaches that do not employ a direct cost function—e.g., principal component regression or partial least squares. Hence, we use a simplification of the solution via a two-step procedure in which the map $\mathbf{b}$ is first constructed naïvely, and then orthogonalized with regard to $\boldsymbol{C}_2$ using a Gram–Schmidt procedure (*SI Appendix*):

$$\boldsymbol{b} - \omega \frac{\boldsymbol{C}_2 \cdot \boldsymbol{b}}{\boldsymbol{C}_2 \cdot \boldsymbol{C}_2} \boldsymbol{C}_2. \qquad [3]$$

Improving Neural Predictor Discriminant Validity. Ideally, a successful method should remove any correlation with nondeceptive behaviors while retaining as much of its predictive power in the task of interest as possible—i.e., shifting the naïve predictor away from the "predictive confounded" quadrant and toward the "predictive unconfounded" region (Fig. 2*D*). For the region-removal method, using a cutoff *P*-value of 0.05 to mask out voxels that were significantly correlated with the control task led to negligible change in performance as compared to the naïve approach (Fig. 2 *F* and *G*). A more systematic examination of different cutoff values (P < 0.1, 0.2, …, 0.99) further showed that while confounding signals were indeed removed as more regions were masked out, this improvement was largely achieved at the expense of accuracy in the task of interest (Fig. 2*F*). Similarly, the relabeling approach and the regress-out methods were only able to remove confounding signals at the expense of reduced accuracy (Fig. 2 *F* and *G*).

In contrast to the poor to mixed performance of the previous three approaches, we found that the dual-goal tuning approach significantly reduced overgeneralization compared to the naive method (P < 0.001) and was in fact nearly able to eliminate it completely (r = −0.0067, P = 0.85) while retaining predictive power for deceptive choices (single-trial prediction: 56.0%, P = 0.01; Fig. 2 *F* and *G*). This difference reflects an important distinction between the other approaches and dual-goal tuning, in that the latter required much less tradeoff between predictive performance in the two tasks.

Distinguishing between Deceptive and Selfish Behavior. A potentially important limitation of inferring discriminant validity based on the removal or absence of overgeneralization is that it relies on accepting the null hypothesis. Such conclusions are known to be problematic, as the null may fail to be rejected simply because the study lacked statistical power, for example (37). To address this possibility, we constructed a positive test involving a "high confound" testing set consisting of deceptive and selfish trials.

Here, a helpful analogy can be made with pregnancy tests, which when used in the general population are known to be highly sensitive and specific. However, because the test does not measure pregnancy per se, but levels of β-hCG hormone, the test is known to perform poorly if used in a "high-confound" testing set in which pregnant women are intermixed with populations with syndromes causing abnormally high levels of β-hCG, such as trophoblastic disease and certain cancers (38). Thus, just as a pregnancy test with improved discriminant validity can be demonstrated by successfully distinguishing between pregnant and other individuals with elevated β-hCG levels (38), we can provide positive evidence that discriminant validity has improved by showing that the corrected neural predictor can distinguish between deceptive and merely selfish behavior (Fig. 3*A*).

We found that whereas dual-goal tuning was able to significantly distinguish between the two trial types (mean AUC = 53.3%, P = 0.0177), all other methods did not result in greater than chance rates of performance (Fig. 3 *B* and *C*). Furthermore, dual-goal tuning had significantly higher performance than each of the other four methods (P < 0.05 for all tests). Thus, consistent with the indirect evidence provided by our overgeneralization results above, these data provide positive evidence that dual-goal tuning was able to improve discriminant validity of the deception predictor.

Neural Systems Underlying Deception. Beyond prediction, the ability to detect and control for nuisance processes can also enrich our understanding of neural systems underlying deception. For example, previous meta-analyses of fMRI studies of deception have suggested the involvement of a network of regions in deception, including the anterior insula, anterior cingulate, inferior frontal gyrus, inferior parietal lobule, and superior frontal gyrus (17, 39, 40). However, it is unclear the extent to which predictors based on these regions are vulnerable to the presence of confounding processes captured by our control task.

Using our deception task and searchlight MVPA (41), we corroborate previous meta-analytic findings (17) and show that regions such as the superior frontal gyrus and precuneus contain signals that allow us to decode deception (Fig. 4*A*). We used a searchlight with a radius of 2 voxels (33 voxels in a spherical ROI) and a partial least squares (PLS) algorithm with leave-one-subject-out cross-validation, followed by whole-brain permutation testing of significant predictive performance. We also find evidence that overgeneralization at the ROI level is significantly more likely to occur than would be expected by chance such that predictors trained on the deception task also generalize to the control task (permutation test P = 0.003; Fig. 4*B*). Contingent on the stringency of the null criterion for meaningful overgeneralization, a more systematic examination of different cutoff values further shows that the percentage of voxels that overgeneralize ranges between 18 and 88% (P < 0.05: 18%, P < 0.1: 28%, P < 0.2: 39%, P < 0.5: 66%, P < 0.8: 88%).

Importantly, the extent to which the predictive regions were affected by posttraining orthogonalization sheds light on the nature of neural signals in these regions. Some highly predictive regions, such as the left occipital pole, were no longer able to significantly predict deception after dual-goal tuning, suggesting that their predictive power was entirely driven by confounded signals (Fig. 4*C*). In contrast, other highly predictive regions, such as the superior frontal gyrus, were able to retain predictive power even after orthogonalization, suggesting the comingling of distinct signals within these regions (details and coordinates in *SI Appendix*, Table S2).

Discussion

Deception is ubiquitous in nature and a known feature of a number of mental and behavioral disorders (42, 43). Despite its importance, however, studies on the relationship between deception and the underlying biological mechanisms have long been discounted because of a historical lack of attention to its scientific foundations. As the National Research Council lamented in its 2003 report on the poor knowledge of diagnostic and psychometric properties of lie detection techniques, "More intensive efforts to develop the basic science in the 1920s would have produced a more favorable assessment in the 1950s; more intensive efforts in the 1950s would have produced a more favorable assessment in the 1980s; more intensive efforts in the 1980s would have produced a more favorable assessment now" (4).

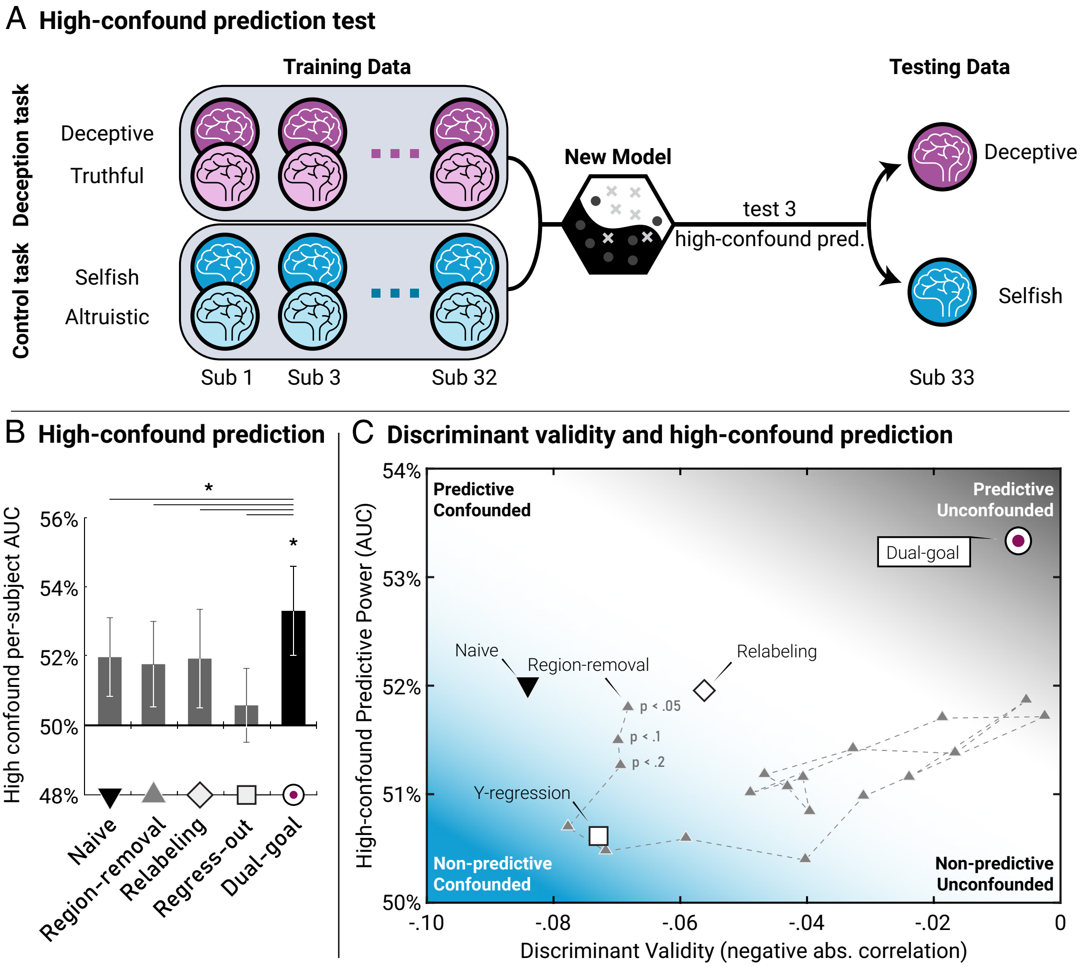


Fig. 3. (*A*) Schematic of test 3: prediction in a high-confound test set in which lie trials are mixed with a confounded selfish but nondeceptive behavior. (*B*) Bar graphs of validity tests and high-confound prediction performances. Only dual-goal tuning significantly discriminated between lie trials and selfish trials. Dual-goal tuning also showed significantly higher AUC than each of the other four methods at $P < 0.05$. (*C*) Discriminant validity (test 2) and high-confound prediction performances (test 3) of different neural predictors compared on a 2D plot. In contrast with dual-goal tuning, all other methods fail to adequately control nuisance correlation and hence result in low discriminability between highly confounded trials. *$P < 0.05$.

Our work seeks to break this stalemate by providing such a scientific foundation. First, we build upon pioneering cognitive neuroscience studies in which participants could choose to lie, rather than being instructed to do so (8, 10, 14, 44–46). By capturing the fact that honesty and deception are properties of the communicative signals that agents send to one another in the service of some economic or evolutionary benefit, meta-analyses of fMRI studies using these paradigms have identified a consistent set of brain regions in the lateral and medial PFC, as well as the anterior insula, that are more strongly engaged by deceptive compared with honest responses (17, 47). More recent efforts have extended these findings by incorporating machine learning methods to predict behavior from brain activity, permitting researchers to avoid problems associated with reverse inference—e.g., that connectivity patterns in the self-referential thinking network were able to predict the honesty of participants during decision-making (14).

Building on these efforts, we sought to explore a complementary aspect of prediction and generalization: the need to evaluate for the presence or absence of potentially confounding processes. Specifically, we leveraged generalization tests in MVPA methods to test for the presence of confounding processes in a deception predictor. Generalization tests have become increasingly popular in cognitive neuroscience (12, 25–30, 48, 49), and they can be used to assess the validity of the predictive signals. Our finding that signals underlying the neural predictor of deception are associated with nondeceptive, selfish choices supports long-standing validity concerns that predictive signals for deception can be driven by confounding processes.

In addition, our results suggest that using cross-task generalization to identify confounding signals can provide essential information about construct validity in a manner that extends beyond within-dataset sensitivity and specificity (12, 17). To clearly disambiguate these distinct contributions, we note that measures such as sensitivity or specificity address criterion validity using metrics such as the percentage of the total number of lies the predictor identifies, or the percentage of truths it falsely flags as lies. Cross-task generalization, on the other hand, probes the construct validity of the predictive signal by asking whether the predictive signal is correlated with unrelated measures, such as being significantly greater for certain types of truth trials (selfish ones) than other truth trials (altruistic ones) (18).

Our second contribution is to develop an approach to create predictors that do not demonstrate undesirable out-of-sample generalization when applied to a new task. While there have been studies in which an absence of generalization across two tasks has been used to show evidence for the distinctiveness of two constructs (e.g., ref. 30), developing methods to eliminate an already existing generalization has received less attention. Rather, there were efforts to test whether the level of prediction is beyond what is expected from confounds (50, 51). Such tests can be used in

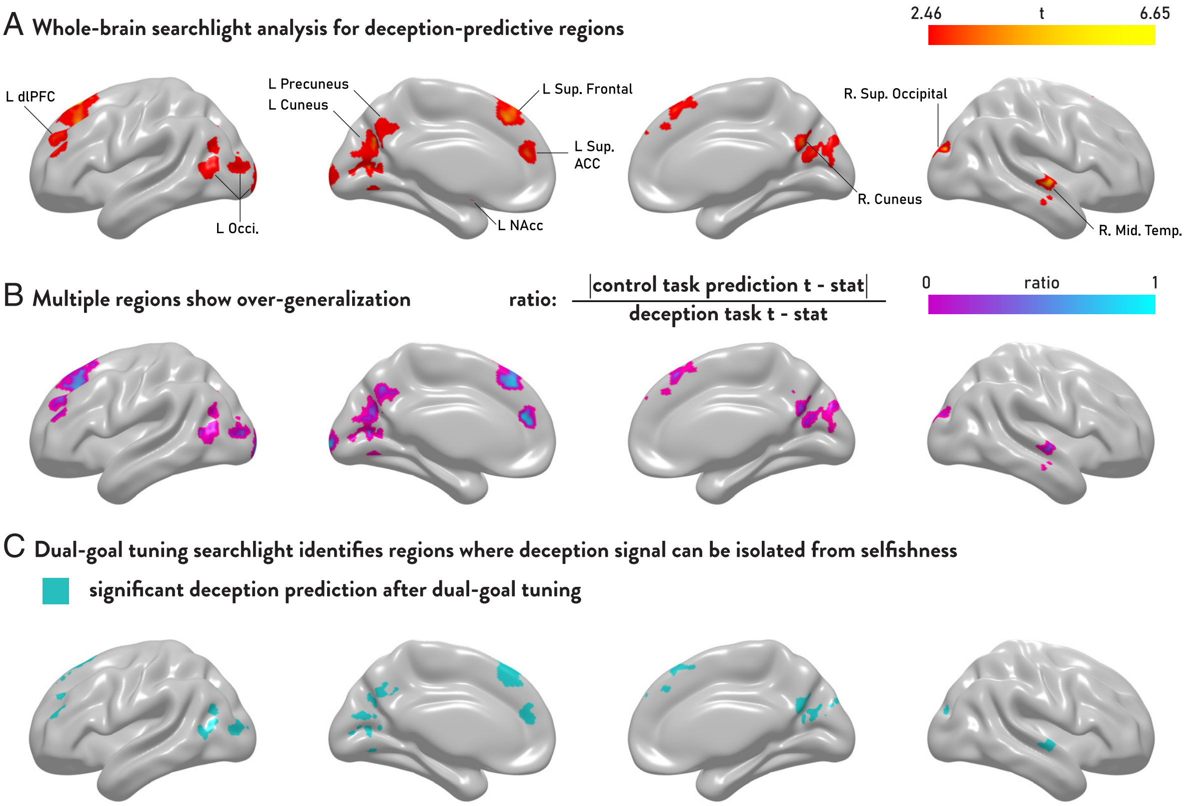

A Whole-brain searchlight analysis for deception-predictive regions
2.46
t
6.65
L dlPFC
L Precuneus
L Cuneus
L Sup. Frontal
L Sup. ACC
L Occi.
L NAcc
R. Sup. Occipital
R. Cuneus
R. Mid. Temp.
B Multiple regions show over-generalization
ratio: |control task prediction t - stat| / deception task t - stat
0
ratio
1
C Dual-goal tuning searchlight identifies regions where deception signal can be isolated from selfishness
significant deception prediction after dual-goal tuning

Fig. 4. (*A*) Searchlight analysis for regions that can predict deceptive choices. Searchlight analysis with a radius of two voxels was performed across the entire brain to identify regions that significantly predict deceptive choices, as assessed by leave-one-out cross-validation. Regions with predictive performances significantly above 50% at the whole-brain correction level (permutation tested TFCE $P < 0.05$) are shown. (*B*) Cross-task generalization performance is measured for regions identified in (*A*); the ratio of the *t* test statistics is shown. Several regions that have high predictive power in panel (*A*) are also shown to have high generalization in panel (*B*). (*C*) Dual goal tuning is applied at each searchlight to eliminate cross-task generalization and thereby identify regions that can significantly predict deceptive but not selfish choices ($P < 0.05$).

cases where a confounding variable is comeasured with the targeted behavior, but is difficult to use in the generalized case here where there are separate tasks for signal vs. confound. Furthermore, while pioneering work on regression models and machine learning algorithms in neuroimaging has primarily addressed the goal of tuning model hyperparameters to improve predictive performance (28, 52), purging confounding signals will in general require sacrificing, rather than improving, performance.

Our results suggest that controlling for overgeneralization can be achieved by addressing the predictor construction directly rather than altering what is included in the training data. Preprocessing the training data by removing the most confounded voxels (i.e., region-removal) performs poorly when signals of interest and no-interest comingle at the voxel level. In such cases, region-removal can result in reduced power in predicting the variable of interest due to imperfect orthogonalization (Fig. 5*A*). On the other hand, our dual-goal tuning procedure can be seen as a shearing transformation that reduces the inner product between the prediction map and the nuisance covariance (Fig. 5*C*).

In contrast, by introducing data from the control task to the training process without changing the optimization function, the relabeling approach attempts to find a decision boundary to distinguish the lie trials from all other trials. However, while this procedure enriches the examples of "truth" trials in order to identify a better decision threshold, the predictive signal, which runs orthogonal to the decision boundary, is still correlated with the nuisance signal (Fig. 5*E*). On the other hand, by conceptualizing the problem as a form of constrained optimization for out-of-sample prediction, dual-goal tuning orthogonalizes the predictive signal with regard to the nuisance signal such that the predictor cannot distinguish between the selfish and altruistic nondeceptive choices no matter where the decision boundary is placed.

At the neural level, we show that removing the presence of confounding signals can allow for more specific inferences about the nature of underlying processes. Searchlight analysis of deceptive behavior revealed a number of regions that previous studies have implicated in deception, including the anterior cingulate cortex (ACC), superior frontal gyrus (SFG), dorsolateral prefrontal cortex (dlPFC), and nucleus accumbens (NAcc) (8, 14, 17, 47) (*SI Appendix*, Table S2). At the same time, as with the whole-brain predictor, the discriminant validity of these findings is significantly undercut by the widespread presence of signals that are not specific to deception. By purging signals that are not specific to deception, our dual-goal tuning approach shows that in many of these regions, a neural signature of deception that is not shared with the control task can be identified by incorporating control task data into training.

Notably, while traditionally sensory and unimodal areas, such as the occipital pole, no longer predict deception after shared signals are removed, surviving polymodal regions support hypotheses that deceptive behaviors utilize both domain-general and domain-specific higher-order processes linked to cognitive control, self-referential thought, and social cognition (53–55). Importantly, the fact that a neural predictor does not generalize to the control task does not imply the underlying processes are "unique" to deception. Indeed, although the two games are isomorphic in the game-theoretic sense, in that the games differ only in the surface labels, the fact that participants behave quite differently in the two games suggests that these labels can matter a great deal to message

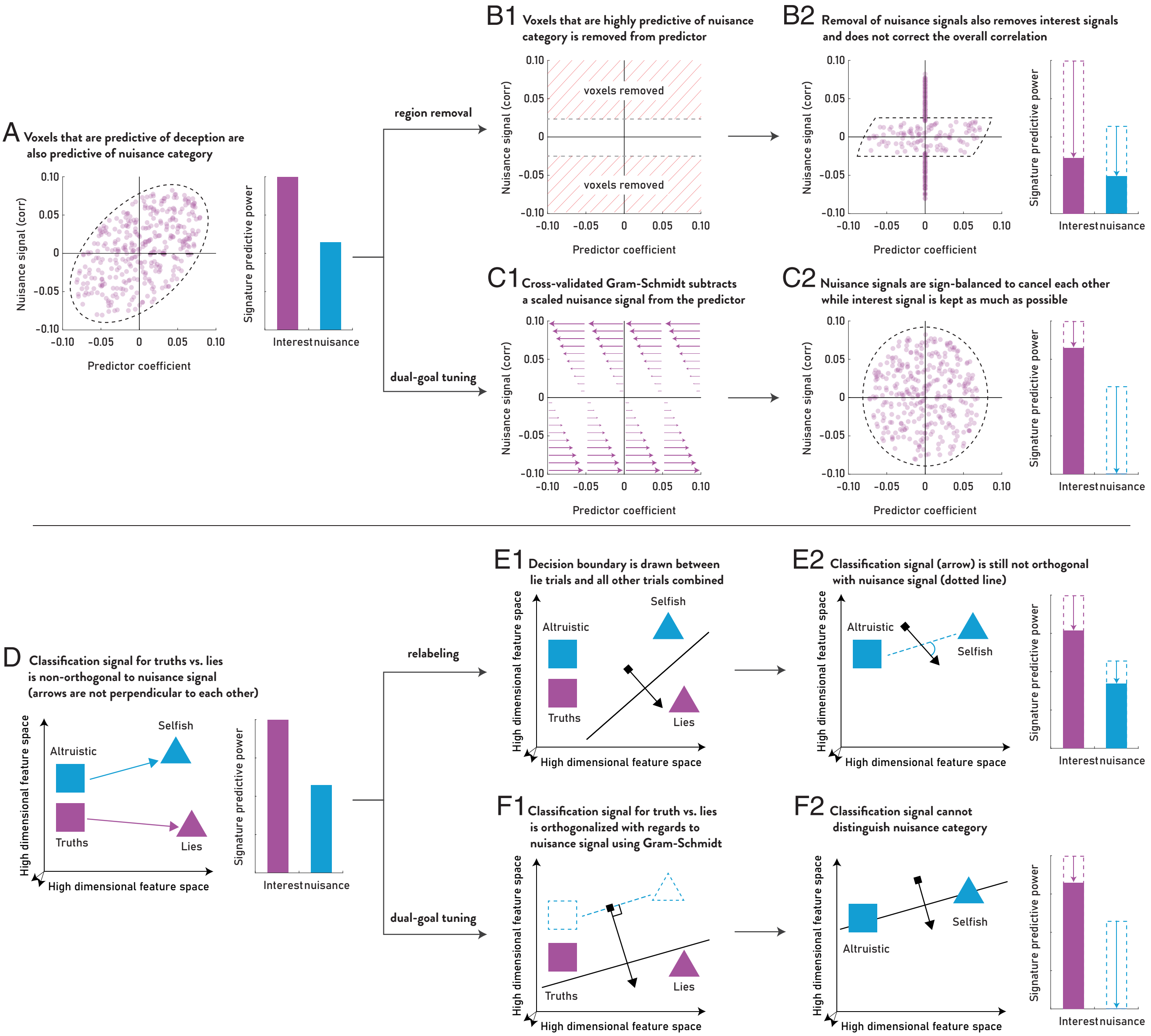

A Voxels that are predictive of deception are also predictive of nuisance category
Nuisance signal (corr)
Predictor coefficient
Signature predictive power
Interest
nuisance
region removal
dual-goal tuning
B1 Voxels that are highly predictive of nuisance category is removed from predictor
voxels removed
voxels removed
B2 Removal of nuisance signals also removes interest signals and does not correct the overall correlation
C1 Cross-validated Gram-Schmidt subtracts a scaled nuisance signal from the predictor
C2 Nuisance signals are sign-balanced to cancel each other while interest signal is kept as much as possible
0.10
0.05
0
−0.05
−0.10
D Classification signal for truths vs. lies is non-orthogonal to nuisance signal (arrows are not perpendicular to each other)
High dimensional feature space
Altruistic
Selfish
Truths
Lies
relabeling
E1 Decision boundary is drawn between lie trials and all other trials combined
E2 Classification signal (arrow) is still not orthogonal with nuisance signal (dotted line)
F1 Classification signal for truth vs. lies is orthogonalized with regards to nuisance signal using Gram-Schmidt
F2 Classification signal cannot distinguish nuisance category

Fig. 5. Comparison of region removal, relabeling, and dual-goal tuning in the presence of confounding processes. (*A*) Depiction of an underlying signal for which the signals of interest are confounded with nuisance signals. Voxels that carry highly positive signal for the signal of interest also carry highly positive signal for the nuisance signal, and vice versa for negative signals. (*B1* and *B2*) Depiction of the region-removal method, in which voxels that are strongly correlated with nuisance signals are removed before building a predictor. However, because the underlying nonorthogonality has not been solved, the region-removal approach is unlikely to achieve orthogonality. Furthermore, as the voxels that are correlated with nuisance are removed, so are the voxels that are correlated with signal of interest. (*C1* and *C2*) Depiction of the dual-goal tuning approach using Gram–Schmidt orthogonalization to correct the predictor. The shearing transformation is controlled by the orthogonalization hyperparameter so as to achieve zero out-of-sample predictive power for nuisance. (*D*) Depiction of a classification space in which the underlying signal of interest (purple arrow) is not orthogonal (perpendicular) to the nuisance signal (blue arrow). (*E1* and *E2*) Depiction of the relabeling approach, in which a best decision boundary is identified between the lie trials (purple triangle) and all other trials. While the decision boundary may be effective in labeling all truth trials as truths, the predictive signal (purple arrow) is still correlated with the nuisance signal such that selfish trials still receive higher prediction scores than altruistic trials. (*F1* and *F2*) Depiction of the dual-goal tuning approach, in which the predictive signal (purple arrow) is constructed under dual-goal tuning to be orthogonal (perpendicular) to the nuisance signal (blue dotted line) such that the predictor cannot reliably distinguish between altruistic and selfish trials.

senders. It is possible, for example, that predictive power in at least some ROIs reflects engagement of processes involved in parsing and evaluating the truth value of messages only in the deception condition. This idea is consistent with past hypotheses that deception may require additional engagement of inhibitory control, working memory, task switching processes, and other executive functions that are subserved by regions including the SFG and ACC found in our study (17, 40, 56–59). This idea is similarly true for social cognitive processes such as those underlying impression management, which may be differentially involved in deception compared to nondeceptive decision (3). The fact that we can control for shared processes therefore raises the exciting possibility that future studies may be able to test these hypotheses in more specific ways, for example by incorporating control conditions that vary in their engagement of executive functioning or social cognitive processes.

More generally, while we advance upon previous paradigms in which participants are instructed *when* to lie, our signaling task and others involving motivated deception are still limited by the fact that experimenters specify *how* participants can lie. Addressing this issue will require different behavioral paradigms and models that allow researchers to better capture the open-ended nature

(60) of real-world deception. In turn, such paradigms might be used to elucidate neurocognitive processes that underlie decisions regarding when, how, and whom to deceive. Additional work is also needed to account for individual-level heterogeneity in out-of-sample prediction, which is critical in forensic and clinical settings (17, 40, 56–59). Depending on the nature of the questions, one can either treat individual heterogeneity as a factor of no interest—e.g., in order to identify core (shared) components of deception—or more fully characterize the neural heterogeneity of deception in order to account for it when making out-of-sample predictions.

Finally, we note that although discriminant validity may be particularly important in the case of deception, it is also of critical importance in other areas of cognitive neuroscience and computational psychiatry (11, 12, 61). Currently, dissociating co-occurring or confounded processes requires an experiment that allows for orthogonal control of both processes, an approach that is not always possible. Failing that, studies have argued for the distinctiveness of mental processes by showing an absence of overgeneralization across datasets with naïve predictors (29, 30). However, our results suggest that even when there is a considerable amount of overlap, the underlying mental processes may be distinguishable. Our methodology may therefore be useful in dissociating common co-occurring processes, especially if the orthogonality must be established post hoc or used as a complement to specific task designs, such as cases involving working memory and attention (62), valuation and salience (57), or valence, arousal, and emotions (48). In applied settings such as computational psychiatry, our approach may aid in constructing neural biomarkers specific to a particular diagnosis without relying on covariates of no interest (63, 64). Thus, despite important limitations, our findings represent a meaningful advance given the potential significance of the questions and the protracted nature of the challenges involved (4).

Materials and Methods

Participants. Forty healthy individuals provided informed consent and participated in the experiment (16 males; age = Mean 20.8 y ± S.D. 2.6 y). Seven participants were excluded from data analysis because they exhibited one-sided choices in more than 90% of trials in at least one task and therefore lacked sufficient behavioral variation to predict (see below). As a result, 33 participants were included in final data analysis, and no additional participants were removed based on image quality or motion in the scanner. All experimental procedures were approved by the Institutional Review Board of Peking University.

Neural Activity Estimation for Decoding Analyses. Neural prediction was performed at two levels. At the subject level, activity for trials of the same category was estimated together such that one activity image per category was estimated for each run. At the trial-level, each trial's activity was estimated separately. For subject-level predictions, we used a GLM with one regressor for all trials in which the participant made truthful/altruistic choices, and one for deceptive/selfish choices. For single-trial predictions, the GLM consisted of a separate regressor for each trial [beta-series regression (65)]. Regressors were modeled with an impulse function time-locked to the button press and convolved with a double-gamma HRF (see *SI Appendix* for decoding from trial onset). Nuisance regressors included the average CSF activity, the average white matter activity, and the top 10 PCA components derived from the combined CSF+white-matter masks (a_comp_cor, provided by fMRIPrep), as well as 24 motion parameters [6 affine transformations for each TR and the previous TR, and their squares (66)]. The z-statistic of each regressor's coefficient estimate was used as input for further analysis (67).

Out-of-Sample Prediction. Predictions were made using leave-one-subject-out cross-validation in which a predictor was trained on behavior of 32 participants' deception task using the Thresholded Partial Least Squares (T-PLS) algorithm (28), then used to predict the left-out participants' behavior in both deception and control tasks. This procedure was repeated 33 times for all participants (see *SI Appendix* for out-of-sample prediction parameters).

Data, Materials, and Software Availability. The experimental dataset used in this study has been anonymized and is available online at OpenNeuro (DOI: 10.18112/openneuro.ds005128.v1.0.0) (68). All other data are included in the manuscript and/or supporting information.

ACKNOWLEDGMENTS. This study was supported by grant funding from STI2030-Major Projects (2022ZD0205100 to L.Z.), NSFC (32071095, 32325023, T2421004 to L.Z.), the National Institute of Mental Health (R01 MH112775 to M.H. and A.K.), the National Institute on Alcohol Abuse & Alcoholism (R01 AA026587 to A.K. and M.H.), and the NSF (1851902 to M.H.).

Author affiliations: [a]Helen Wills Neuroscience Institute, University of California, Berkeley, CA 94720; [b]School of Psychological and Cognitive Sciences and Beijing Key Laboratory of Behavior and Mental Health, International Data Group/McGovern Institute for Brain Research, Peking-Tsinghua Center for Life Sciences, Peking University, Beijing 100871, China; [c]Department of Neurology, University of California, San Francisco, CA 94158; [d]Division of Neurology, San Francisco Veterans Affairs Health Care System, San Francisco, CA 94121; and [e]Haas School of Business, University of California, Berkeley, CA 94720

1. K. Alder, *The Lie Detectors: The History of an American Obsession* (Simon and Schuster, 2007).
2. B. Kleinmuntz, J. J. Szucko, Lie detection in ancient and modern times: A call for contemporary scientific study. *Am. Psychol.* **39**, 766 (1984).
3. K. E. Sip, A. Roepstorff, W. McGregor, C. D. Frith, Detecting deception: The scope and limits. *Trends Cogn. Sci.* **12**, 48–53 (2008).
4. National Research Council Division of Behavioral Committee on National Statistics Board on Behavioral Sensory Sciences & Committee to Review the Scientific Evidence on the Polygraph, *The Polygraph and Lie Detection* (National Academies Press, 2003).
5. W. M. Marston, Systolic blood pressure symptoms of deception. *J. Exp. Psychol.* **2**, 117 (1917).
6. U.S. Office of Technology Assessment, "Scientific validity of polygraph testing: A research review and evaluation, a technical memorandum" (Report No. OTA-TM-H-15, U.S. Office of Technology Assessment, Washington, DC, 1983).
7. W. A. Searcy, S. Nowicki, *The Evolution of Animal Communication: Reliability and Deception in Signaling Systems: Reliability and Deception in Signaling Systems* (Princeton University Press, 2010).
8. L. Zhu *et al.*, Damage to dorsolateral prefrontal cortex affects tradeoffs between honesty and self-interest. *Nat. Neurosci.* **17**, 1319–1321 (2014).
9. U. Gneezy, Deception: The role of consequences. *Am. Econ. Rev.* **95**, 384–394 (2005).
10. A. C. Jenkins, L. Zhu, M. Hsu, Cognitive neuroscience of honesty and deception: A signaling framework. *Curr. Opin. Behav. Sci.* **11**, 130–137 (2016).
11. R. A. Poldrack, Inferring mental states from neuroimaging data: From reverse inference to large-scale decoding. *Neuron* **72**, 692–697 (2011).
12. P. A. Kragel, L. Koban, L. F. Barrett, T. D. Wager, Representation, pattern information, and brain signatures: From neurons to neuroimaging. *Neuron* **99**, 257–273 (2018).
13. C. Davatzikos *et al.*, Classifying spatial patterns of brain activity with machine learning methods: Application to lie detection. *Neuroimage* **28**, 663–668 (2005).
14. S. P. H. Speer, A. Smidts, M. A. S. Boksem, Cognitive control increases honesty in cheaters but cheating in those who are honest. *Proc. Natl. Acad. Sci. U.S.A.* **117**, 19080–19091 (2020).
15. J.-D. Haynes, Detecting deception from neuroimaging signals–A data-driven perspective. *Trends Cogn. Sci.* **12**, 126–127 (2008).
16. Y. Feng, S. Hung, P. Hsieh, Detecting spontaneous deception in the brain. *Hum. Brain Mapp.* **43**, 3257–3269 (2022).
17. M. J. Farah, J. B. Hutchinson, E. A. Phelps, A. D. Wagner, Functional MRI-based lie detection: Scientific and societal challenges. *Nat. Rev. Neurosci.* **15**, 123–131 (2014).
18. L. J. Cronbach, P. E. Meehl, Construct validity in psychological tests. *Psychol. Bull.* **52**, 281 (1955).
19. C. F. Camerer, *Behavioral Game Theory: Experiments in Strategic Interaction* (Princeton University Press, 2011).
20. S. A. Spence *et al.*, Behavioural and functional anatomical correlates of deception in humans. *Neuroreport* **12**, 2849–2853 (2001).
21. D. D. Langleben *et al.*, Brain activity during simulated deception: An event-related functional magnetic resonance study. *Neuroimage* **15**, 727–732 (2002).
22. T. M. C. Lee *et al.*, Lie detection by functional magnetic resonance imaging. *Hum. Brain Mapp.* **15**, 157–164 (2002).
23. J. G. Hakun *et al.*, "fMRI investigation of the cognitive structure of the concealed information test" in *Neuroscience and Crime*, H. J. Markowitsch, Ed. (Psychology Press, 2020), pp. 59–67.
24. F. A. Kozel *et al.*, Detecting deception using functional magnetic resonance imaging. *Biol. Psychiatry* **58**, 605–613 (2005).
25. L. J. Chang, P. J. Gianaros, S. B. Manuck, A. Krishnan, T. D. Wager, A sensitive and specific neural signature for picture-induced negative affect. *PLoS Biol.* **13**, 1–28 (2015).
26. T. D. Wager *et al.*, An fMRI-based neurologic signature of physical pain. *N. Engl. J. Med.* **368**, 1388–1397 (2013), 10.1056/NEJMoa1204471.
27. S. Lee, J. W. Kable, Simple but robust improvement in multivoxel pattern classification. *PLoS One* **13**, e0207083 (2018).
28. S. Lee, E. T. Bradlow, J. W. Kable, Fast construction of interpretable whole-brain decoders. *Cell Rep. Methods* **2**, 100227 (2022).

29. S. Lee *et al.*, A neural signature of the vividness of prospective thought is modulated by temporal proximity during intertemporal decision making. *Proc. Natl. Acad. Sci. U.S.A.* **119**, e2214072119 (2022).
30. C.-W. Woo *et al.*, Separate neural representations for physical pain and social rejection. *Nat. Commun.* **5**, 5380 (2014).
31. D. T. Campbell, D. W. Fiske, Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychol. Bull.* **56**, 81 (1959).
32. J. Read, B. Pfahringer, G. Holmes, E. Frank, Classifier chains for multi-label classification. *Mach. Learn.* **85**, 333–359 (2011).
33. W. J. Murdoch, C. Singh, K. Kumbier, R. Abbasi-Asl, B. Yu, Definitions, methods, and applications in interpretable machine learning. *Proc. Natl. Acad. Sci. U.S.A.* **116**, 22071–22080 (2019).
34. A. S. Ross, M. C. Hughes, F. Doshi-Velez, Right for the right reasons: Training differentiable models by constraining their explanations. arXiv [Preprint] (2017). https://doi.org/10.48550/arXiv.1703.03717 (Accessed 15 September 2024).
35. C.-L. Hwang, A. S. M. Masud, *Multiple Objective Decision Making–Methods and Applications: A State-of-the-Art Survey* (Springer Science & Business Media, 2012).
36. M. Gori, A. Betti, S. Melacci, *Machine Learning: A Constraint-Based Approach* (Elsevier, 2023).
37. A. N. Kluger, J. Tikochinsky, The error of accepting the "theoretical" null hypothesis: The rise, fall, and resurrection of commonsense hypotheses in psychology. *Psychol. Bull.* **127**, 408 (2001).
38. S. I. McCash, D. J. Goldfrank, M. S. Pessin, L. V. Ramanathan, Reducing false-positive pregnancy test results in patients with cancer. *Obstet. Gynecol.* **130**, 825–829 (2017).
39. M. Delgado-Herrera, A. Reyes-Aguilar, M. Giordano, What deception tasks used in the lab really do: Systematic review and meta-analysis of ecological validity of fMRI deception tasks. *Neuroscience* **468**, 88–109 (2021).
40. N. Lisofsky, P. Kazzer, H. R. Heekeren, K. Prehn, Investigating socio-cognitive processes in deception: A quantitative meta-analysis of neuroimaging studies. *Neuropsychologia* **61**, 113–122 (2014).
41. N. Kriegeskorte, R. Goebel, P. Bandettini, Information-based functional brain mapping. *Proc. Natl. Acad. Sci. U.S.A.* **103**, 3863–3868 (2006), 10.1073/pnas.0600244103.
42. M. J. Vitacco, Syndromes associated with deception. *Clin. Assessm. Maling. Decept.* **3**, 39–50 (2008).
43. A. M. O'Connor, A. D. Evans, Dishonesty during a pandemic: The concealment of COVID-19 information. *J. Health Psychol.* **27**, 236–245 (2022).
44. J. D. Greene, J. M. Paxton, Patterns of neural activity associated with honest and dishonest moral decisions. *Proc. Natl. Acad. Sci. U.S.A.* **106**, 12506–12511 (2009).
45. N. Abe, J. D. Greene, Response to anticipated reward in the nucleus accumbens predicts behavior in an independent test of honesty. *J. Neurosci.* **34**, 10564–10572 (2014).
46. K. E. Sip *et al.*, The production and detection of deception in an interactive game. *Neuropsychologia* **48**, 3619–3626 (2010).
47. S. K. Meier, K. L. Ray, J. C. Mastan, S. R. Salvage, D. A. Robin, Meta-analytic connectivity modelling of deception-related brain regions. *PLoS One* **16**, e0248909 (2021).
48. P. A. Kragel, K. S. LaBar, Multivariate neural biomarkers of emotional states are categorically distinct. *Soc. Cogn. Affect. Neurosci.* **10**, 1437–1448 (2014), 10.1093/scan/nsv032.
49. P. A. Kragel *et al.*, Generalizable representations of pain, cognitive control, and negative emotion in medial frontal cortex. *Nat. Neurosci.* **21**, 283–289 (2018).
50. T. Spisak, Statistical quantification of confounding bias in machine learning models. *Gigascience* **11**, giac082 (2022).
51. D. Chyzhyk, G. Varoquaux, M. Milham, B. Thirion, How to remove or control confounds in predictive models, with applications to brain biomarkers. *Gigascience* **11**, giac014 (2022).
52. L. Grosenick, B. Klingenberg, K. Katovich, B. Knutson, J. E. Taylor, Interpretable whole-brain prediction analysis with GraphNet. *Neuroimage* **72**, 304–321 (2013).
53. V. Menon, M. D'Esposito, The role of PFC networks in cognitive control and executive function. *Neuropsychopharmacology* **47**, 90–103 (2022).
54. M. E. Raichle, The brain's default mode network. *Annu. Rev. Neurosci.* **38**, 433–447 (2015).
55. C. D. Frith, U. Frith, Mechanisms of social cognition. *Annu. Rev. Psychol.* **63**, 287–313 (2012).
56. R. Johnson Jr., J. Barnhardt, J. Zhu, The contribution of executive processes to deceptive responding. *Neuropsychologia* **42**, 878–901 (2004).
57. S. A. Spence, C. Kaylor-Hughes, T. F. D. Farrow, I. D. Wilkinson, Speaking of secrets and lies: The contribution of ventrolateral prefrontal cortex to vocal deception. *Neuroimage* **40**, 1411–1418 (2008).
58. S. E. Christ, D. C. Van Essen, J. M. Watson, L. E. Brubaker, K. B. McDermott, The contributions of prefrontal cortex and executive control to deception: Evidence from activation likelihood estimate meta-analyses. *Cereb. Cortex* **19**, 1557–1566 (2009).
59. D. D. Langleben, Detection of deception with fMRI: Are we there yet? *Legal Criminol. Psychol.* **13**, 1–9 (2008).
60. Z. Zhang *et al.*, Retrieval-constrained valuation: Toward prediction of open-ended decisions. *Proc. Natl. Acad. Sci. U.S.A.* **118**, e2022685118 (2021).
61. T. Davis, R. A. Poldrack, Measuring neural representations with fMRI: practices and pitfalls. *Ann. N. Y. Acad. Sci.* **1296**, 108–134 (2013).
62. K. Oberauer, Working memory and attention–A conceptual analysis and review. *J. Cogn.* **2**, 36 (2019).
63. P. R. Montague, R. J. Dolan, K. J. Friston, P. Dayan, Computational psychiatry. *Trends Cogn. Sci.* **16**, 72–80 (2012), 10.1016/j.tics.2011.11.018.
64. P. Karvelis, M. P. Paulus, A. O. Diaconescu, Individual differences in computational psychiatry: A review of current challenges. *Neurosci. Biobehav. Rev.* **148**, 105137 (2023).
65. J. Rissman, A. Gazzaley, M. D'Esposito, Measuring functional connectivity during distinct stages of a cognitive task. *Neuroimage* **23**, 752–763 (2004), 10.1016/j.neuroimage.2004.06.035.
66. T. D. Satterthwaite *et al.*, Neuroimaging of the Philadelphia neurodevelopmental cohort. *Neuroimage* **86**, 544–553 (2014).
67. M. Misaki, Y. Kim, P. A. Bandettini, N. Kriegeskorte, Comparison of multivariate classifiers and response normalizations for pattern-information fMRI. *Neuroimage* **53**, 103–118 (2010).
68. S. Lee, R. Niu, L. Zhu, A. S. Kayser, M. Hsu, Data from "Deception and signaling task." OpenNeuro. https://dx.doi.org/10.18112/openneuro.ds005128.v1.0.0. Deposited 6 May 2024.

Supporting Information

Image acquisition, preprocessing

MRI images were collected on a Siemens 3T Prisma scanner with a 32-channel head coil. High-resolution T1-weighted anatomical images were acquired using an MPRAGE sequence [repetition time (TR) = 2,530ms; echo time (TE) = 2.98 ms, 512 axial slices, 1 x 0.5 x 0.5mm voxels; 192 x 448 matrix). T2*-weighted functional images were acquired using an EPI sequence with 32 axial slices (3 x 3 x 4.4mm voxels, 64 x 64 matrix, TR = 2,000 ms, TE = 30 ms). All images were preprocessed using fMRIPrep 20.2.1 (1). EPI sequences were skull-stripped, co-registered using boundary-based registration with nine degrees of freedom, head-motion corrected via six degrees of freedom, slice-time corrected, and normalized to a 4mm MNI space. Before the activity estimation procedure below was run, the images were smoothed with a FWHM 8mm Gaussian kernel. The size of the kernel was motivated by prior whole-brain MVPA studies using T-PLS (2, 3) as well as research examining the effects of spatial smoothing on multivariate patterns (4).

Regress-out methods

Regression is widely used to orthogonalize two vectors. In the current study, we have two datasets {$\boldsymbol{X}_1$, $\boldsymbol{Y}_1$}, and {$\boldsymbol{X}_2$, $\boldsymbol{Y}_2$}, where $\boldsymbol{X}$ is a matrix of brain images (with rows comprised of trials and columns consisting of voxels) and $\boldsymbol{Y}$ is the behavior (a single column vector of binary choices). In this situation, one could potentially use regression in multiple ways. However, it is important to first note that, from a practical point of view, all regress-out approaches listed here are less versatile than other methods discussed in the manuscript. The regress-out methods require two datasets to be of the exact same size and to have a coupled relationship between observations. In other words, all observations of $\boldsymbol{Y}_2$ need to be matched one-to-one with all observations of $\boldsymbol{Y}_1$ in order for the regression to work. In the case of the current study, because the two tasks are isomorphic and have trials with the same payoff structure, a reasonable pairing can be constructed to enable trial-level prediction. However, if one considers a more general study in which one wishes to rule out confounding signals generated by other tasks (e.g. a stroop task, a delay discounting task, or others), it becomes impossible to identify which trials of the two unrelated tasks best correspond to each other.

The 'Y-Y regress-out'. Perhaps the most intuitive of all regress-out methods is to regress behavior $\boldsymbol{Y}_2$ (i.e. selfishness) out of behavior $\boldsymbol{Y}_1$ (i.e., deception), with the hopes that removing the behavioral co-variation also results in neurally orthogonal predictors. Conceptually, the 'Y-Y regress-out' approach assumes that the *only* reason for over-generalization is the similarity in behavior between the $\boldsymbol{Y}$s, and that the underlying neural processes (i.e. the $\boldsymbol{X}$s) are independent. As shown in the main manuscript, this regress-out method did not successfully remove overgeneralization, suggesting that the underlying neural representations include shared components. As a formal example for why this method might frequently fail, one could conceive of two uncorrelated behaviors $\boldsymbol{Y}_1$ and $\boldsymbol{Y}_2$ with a corresponding neural signal $\boldsymbol{X}$ that reflects a sum of $\boldsymbol{Y}_1$ and $\boldsymbol{Y}_2$. In such a case, regardless of the orthogonality of $\boldsymbol{Y}_1$ and $\boldsymbol{Y}_2$, a predictor based on $\boldsymbol{X}$ will predict both $\boldsymbol{Y}_1$ and $\boldsymbol{Y}_2$.

The failure of this method in the current manuscript also has implications for creating orthogonal neural predictors (i.e., predictors that don't over-generalize) through factorial experimental designs. For example, Lee et al. (34) used a two-by-two factorial experimental design in which participants were asked to imagine scenarios that were either vivid/non-vivid with positive/negative valence. From this balanced design, they were able to create a neural predictor of imagination vividness that did not predict valence, as well as a neural predictor of imagination valence that did not predict vividness. While the orthogonality of the ***Y***s (vividness and valence) was sufficient to result in orthogonal predictors in their case, it was also possible that orthogonality could not be achieved had there been a number of brain regions whose signal was influenced by both vividness and valence.

The 'X-Y regress-out'. Another possible approach is to remove from $\boldsymbol{X}_1$ (i.e., deception brain data) any correlation with $\boldsymbol{Y}_2$ (i.e., selfishness behavior). This choice is motivated by the fact that if all variations related to selfishness are removed from neural data, no linear combination of neural data can produce an outcome that is correlated with selfishness. This orthogonality, however, relies on being able to provide only pre-cleaned neural data, which is not tenable in out-of-sample prediction. If a predictor is trained on brain data from which the selfish signal has been removed, it will pick up the most predictive voxels without knowing that many of those voxels also contained nuisance signals prior to regress-out. Hence, when the time comes to use the predictor in a real test – i.e. data that has not been pre-cleaned – the predictor will utilize voxels that also identify confounding processes.

The 'X-X regress-out'. The last approach is to remove the neural variance in $\boldsymbol{X}_2$ (control task) from $\boldsymbol{X}_1$ (deception task). This method is similar to the aforementioned 'X-Y regress-out' except that instead of regressing out $\boldsymbol{Y}$ from all voxels of $\boldsymbol{X}$, each voxel of $\boldsymbol{X}_2$ would be regressed out from their corresponding voxel in $\boldsymbol{X}_1$. The difficulty with this method is the same as that in 'X-Y regress-out': the out-of-sample data does not come cleaned.

Dual-goal tuning

Given two datasets ($\boldsymbol{X}_1$, $\boldsymbol{Y}_1$) and ($\boldsymbol{X}_2$, $\boldsymbol{Y}_2$), the predictor map ***b,*** and the nuisance covariance map $\boldsymbol{C}_2 = \boldsymbol{X}_2^T\boldsymbol{Y}_2$, consider the following loss function for ridge regression (note, however, that we show the sum-of-squares loss function below so that we can work with its closed-form solution later; for application to classification problems, logistic or probit loss functions may be more appropriate, as shown in the main manuscript):

$$L(\boldsymbol{b},\lambda,\omega) = (\boldsymbol{Y}_1 - \boldsymbol{X}_1\boldsymbol{b})^T(\boldsymbol{Y}_1 - \boldsymbol{X}_1\boldsymbol{b}) + \lambda\boldsymbol{b}^T\boldsymbol{b} + \omega\boldsymbol{b}^T\boldsymbol{C}_2 \qquad (1)$$

While elegant in form, such a formulation of the cost function has two shortcomings. Firstly, it limits the dual-goal tuning approach to cost-function based methods and excludes data-reduction approaches such as partial least squares or principal component regression. Secondly, the computational burden of identifying a constrained optimization solution for out-of-sample prediction can be significant, especially in neural prediction studies. Because the constraint needs to be satisfied with regards to all possible cross-validation results, the search for ω is particularly taxing, as it needs to be precisely tuned for each of the λ considered – approximately an O(n^2) process. In other words, the entire cross-validation needs to be performed numerous times for each given set of hyperparameters λ. While not infeasible, we consider a more simplified approach below that approximates an O(n) process and therefore reduces computation time. Consider the analytical solution for the ridge cost function above:

$$b = (\boldsymbol{X}_1^T\boldsymbol{X}_1 + \lambda\boldsymbol{I})^{-1}\boldsymbol{X}_1^T\boldsymbol{Y}_1 - \omega(\boldsymbol{X}_1^T\boldsymbol{X}_1 + \lambda\boldsymbol{I})^{-1}\boldsymbol{C}_2 \qquad (2)$$

This solution can be understood as the sum of a naïve ridge predictor $(\boldsymbol{X}_1^T\boldsymbol{X}_1 + \lambda\boldsymbol{I})^{-1}\boldsymbol{X}_1^T\boldsymbol{Y}_1$ and a correction term $\omega(\boldsymbol{X}_1^T\boldsymbol{X}_1 + \lambda\boldsymbol{I})^{-1}\boldsymbol{C}_2$. From this formulation, we consider a more explicit approach in which we start with a naïve predictor, then apply an orthogonalizing correction post-hoc. Here we use a Gram-Schmidt approach in which the map b is first constructed naïvely, and then orthogonalized with regards to $\boldsymbol{C}_2$:

$$\boldsymbol{b} - \omega \frac{\boldsymbol{C}_2 \cdot \boldsymbol{b}}{\boldsymbol{C}_2 \cdot \boldsymbol{C}_2} \boldsymbol{C}_2 \tag{3}$$

Importantly, while ω is typically equal to 1 for in-sample orthogonalization via Gram-Schmidt, in our case the orthogonalization will be performed using the training sample's estimate of $\boldsymbol{C}_2$, which will be different from the out-of-sample's true $\boldsymbol{C}_2$. To demonstrate the necessity for the scaling factor ω, let the in-sample's estimate of $\boldsymbol{C}_2$ be denoted as $\widehat{\boldsymbol{C}_2}$. Then, if we orthogonalize our predictor with regards to $\widehat{\boldsymbol{C}_2}$, the out-of-sample dot product between our orthogonalized predictor and true $\boldsymbol{C}_2$ is as follows:

$$\left\{ \boldsymbol{b} - \frac{\widehat{\boldsymbol{C}_2}^T \boldsymbol{b}}{\widehat{\boldsymbol{C}_2}^T \widehat{\boldsymbol{C}_2}} \widehat{\boldsymbol{C}_2} \right\}^T \boldsymbol{C}_2 = \boldsymbol{b}^T \boldsymbol{C}_2 - \frac{\widehat{\boldsymbol{C}_2}^T \boldsymbol{b}}{\widehat{\boldsymbol{C}_2}^T \widehat{\boldsymbol{C}_2}} \widehat{\boldsymbol{C}_2}^T \boldsymbol{C}_2 = \boldsymbol{b}^T \boldsymbol{C}_2 - \boldsymbol{b}^T \widehat{\boldsymbol{C}_2} \frac{\widehat{\boldsymbol{C}_2}^T \boldsymbol{C}_2}{\widehat{\boldsymbol{C}_2}^T \widehat{\boldsymbol{C}_2}} \tag{4}$$

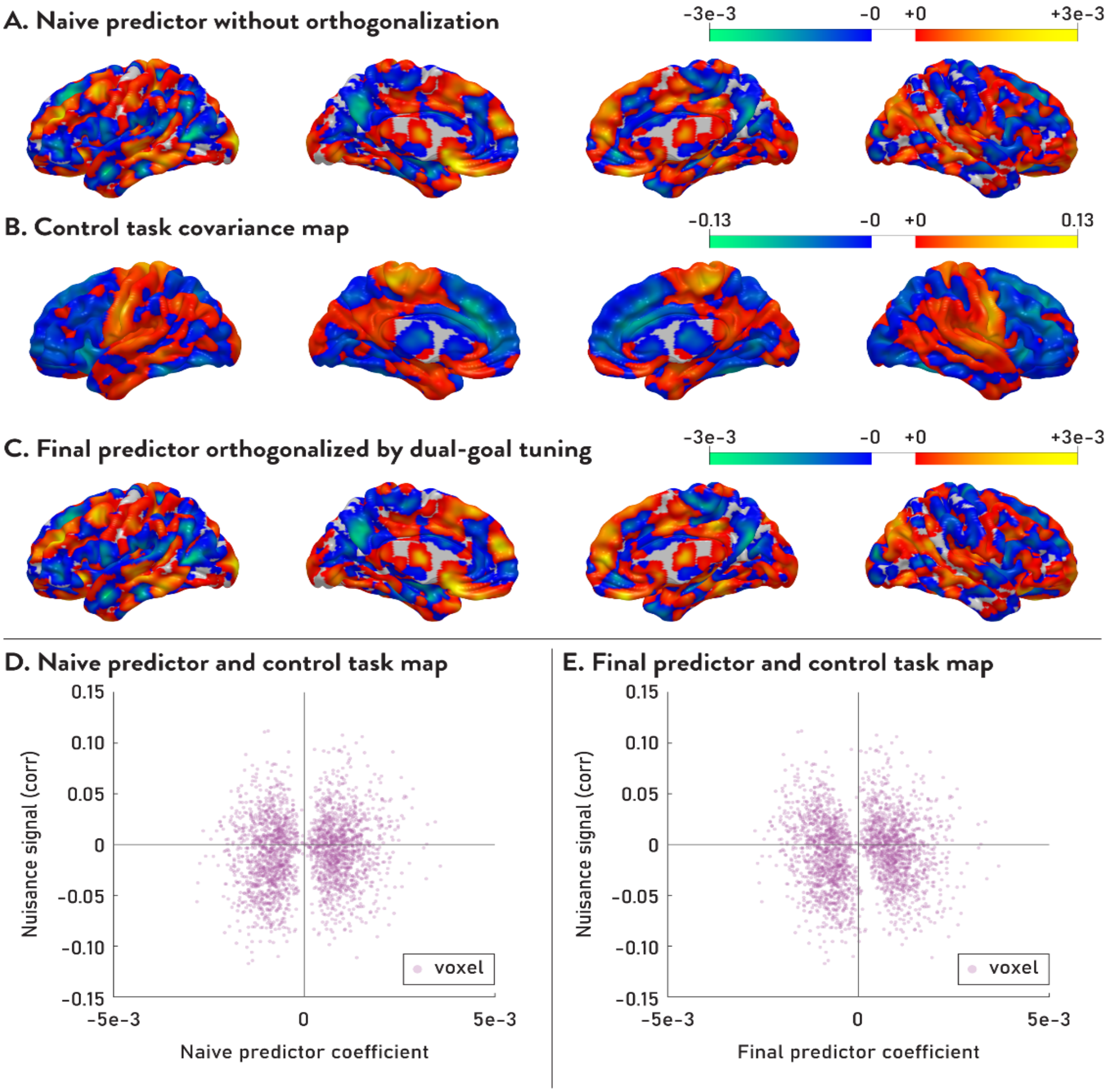
In the last part of the equation, $\boldsymbol{b}^T \boldsymbol{C}_2$ and $\boldsymbol{b}^T \widehat{\boldsymbol{C}_2}$ should be equal to each other in expectation, but the ratio of $\left(\widehat{\boldsymbol{C}_2}^T \boldsymbol{C}_2\right) / \left(\widehat{\boldsymbol{C}_2}^T \widehat{\boldsymbol{C}_2}\right)$ will be less than 1 in expectation given that the denominator is a dot product of the same vector while the numerator is that of different vectors of roughly similar norms (Cauchy-Schwarz inequality). Hence, the orthogonalization will likely be insufficient, which means that the hyperparameter ω should be greater than 1 to compensate for this effect.

Out-of-sample prediction parameters

In both whole-brain prediction analyses and searchlight analyses, we trained the neural predictors using the Thresholded Partial Least Squares (T-PLS) algorithm. T-PLS employs two hyperparameters that require tuning for optimal predictions: the number of PLS components, and the thresholding level. We sought to maintain similar hyperparameter ranges across the whole-brain prediction as well as the searchlight. In whole-brain predictions, we used a maximum of 5 PLS components with thresholding level between 50% to 100%. We did not search for solutions under 50% as too few voxels can make the orthogonalization solution unstable. For searchlight analyses, we held the thresholding at 100% because each searchlight only had 33 voxels, too few to estimate a reasonable cutoff value. In order to tune the hyperparameters of T-PLS, we employed another leave-one-out cross-validation within the training data (i.e., nested cross-validation). Specifically, we trained a T-PLS model on 31 participants and used it to predict the left-out participant in the training data while the hyperparameters were varied.

Whole-brain predictor before and after dual-goal tuning

The final whole-brain predictor of deception, trained from all data using T-PLS and dual-goal tuning, shows interpretably clustered coefficients across the entire brain (**Figure S1**). Strongly positively-weighted regions include the dorsomedial PFC, the dorsolateral PFC, and ventromedial PFC, while the precuneus was notably negatively-weighted. These regions correspond well with the areas identified in the searchlight analysis (**Figure 4**). Notably, orthogonalizing the naive neural predictor (**Figure S1**, panel A) by subtracting the weighted covariance map from the control task (**Figure S1,** panel B) does not substantially change the predictor's final appearance (**Figure S1**, panel C). Even as it implements the Gram-Schmidt orthogonalization, the shearing transformation maintains the general pattern of coefficients, resulting in a final map that retains the interpretability of the naive map.



A. Naive predictor without orthogonalization

B. Control task covariance map

C. Final predictor orthogonalized by dual-goal tuning

D. Naive predictor and control task map

E. Final predictor and control task map

Figure S1. Final whole-brain predictor before and after dual-goal tuning. Panel A shows the naive whole-brain predictor created using the T-PLS algorithm on all 33 participants' data, for visualization purposes. (Note that for the cross-validation analyses described elsewhere, the predictor was constructed using a leave-one-out procedure that included 33 different sets of 32 training subjects and one test subject.) The dual-goal tuning approach subtracts a scaled version of the covariance map derived from the control task (panel B) from the naive map (panel A). The resulting map, shown in panel C, has an overall appearance quite similar to the naive map in panel A. To explain this similarity, panels D and E show, for the naive predictor and the final predictor respectively, the scatterplots of the predictor coefficients (abscissa) and nuisance signal correlation (ordinate) for every voxel in the brain. As reflected in the apparent rotation of the data from panel D to panel E, dual-goal tuning implements a shearing transformation that reduces the size of the predictor coefficients for voxels in the first and the third quadrants, while increasing them for voxels in the second and fourth quadrants, thereby achieving orthogonality as defined by the dot product. The fact that the general pattern of coefficients remains largely unaltered results in visually negligible differences despite a significant change in generalization properties.

Decoding from trial onset versus response time

Given the complexity of the task, we thought that decisions in this task would be better captured by a model that is time-locked to the choice response itself. At the request of a reviewer, we re-analyzed the data with the neural estimates obtained from the trial onset. Somewhat surprisingly, we were unable to significantly predict either task for out-of-sample prediction in either condition (deception mean AUC: 50.58%, p = 0.75, preference mean AUC: 53.17%, p = 0.09). Based on these results, it seems that most of the predictive signals may be time-locked to choice response rather than choice presentation.

Subject-level prediction and validity test performances of various methods

As in the trial-level prediction results shown in the main manuscript (**Figure 2G** & **Figure 3B**), all methods used retained significant predictive performances at the subject-level for the within-task prediction (test 1), but only dual-goal tuning resulted in meaningful elimination of cross-task generalization (test 2). As a result, only dual-goal tuning was able to successfully distinguish lie trials from selfish trials in high-confound prediction (test 3). Regress-out methods could not be included in subject-level prediction as the coupling relationship cannot be established between the observations of two tasks when the observations are grouped by the participants' choices.

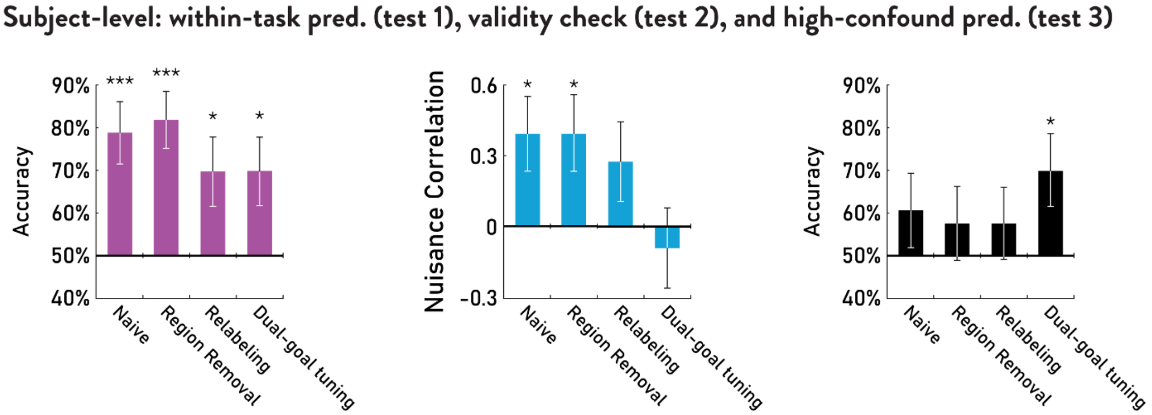


Figure S2. Bar Graphs of subject-level tests for within-task prediction (test 1, left in purple), cross-task validity check (test 2, middle in blue), and high-confound prediction (test 3, right in black). * *p* < .05, *** *p* < .001.

Experimental Design and payment

We used a signaling task in which participants, in the role of the sender, chose one of two messages to send to a recipient (5, 6). On each trial, participants saw two monetary allocations on the screen. One of the monetary allocations provided more money for the participant than the recipient, while the other allocation provided more money for the recipient than the participant. Participants completed two versions of the task, with task order counterbalanced across participants. In the main deception task, participants could choose between two messages: 'option A will earn you more money than option B' and 'option B will earn you more money than option A'. Given the allocations A and B, only one of these two messages was true; the other message was deceptive but yielded more money for the participant. Hence, participants made decisions between sending an altruistic honest message and a deceptive selfish message (**Figure S3**). In the control task, while the monetary allocations were the same, the message that participants could send was either 'I would prefer you to choose option A' or 'I would prefer you to choose option B’. In this case, neither message was deceptive; instead, each was a statement of preference. All participants completed 46 trials for each task, including 6 “non-conflict” trials in which one option brought more money for both the participant and the recipient than the other (see **Table S1** for the allocations presented on individual trials). Within each task, the order of the trials was randomized.

Significantly, for both conditions, participants were aware that as the sender, they could see the monetary options but could not make the choice between them, while the receiver could not see the monetary options but was responsible for the choice between them. This important manipulation renders the receiver completely reliant upon the sender for any information about the choice, preventing the receiver from using payoff information to infer the sender’s behavior. In addition, senders were aware both that in the deception condition, recipients would never know the truth value of the message they received, and that they as the sender were paired with a separate receiver on each trial, limiting influences of any one trial on subsequent trials (5, 6).

Participants were never given feedback from the receivers about their choices. Instead, they were informed that the receiver’s choice was consistent with the sender’s messages 78% of the time, according to previous studies. Participants were told that one trial in each task would be selected for payment at the end of the experiment, consistent with the above expectation that the chosen option would be realized 78% of the time and the unchosen option the remaining 22% of the time. After the comprehension quiz and practice trials in the scanner, the signaling tasks were presented as described (see supplemental slides for task instruction slides, translated into English with appropriate monetary currency conversion).

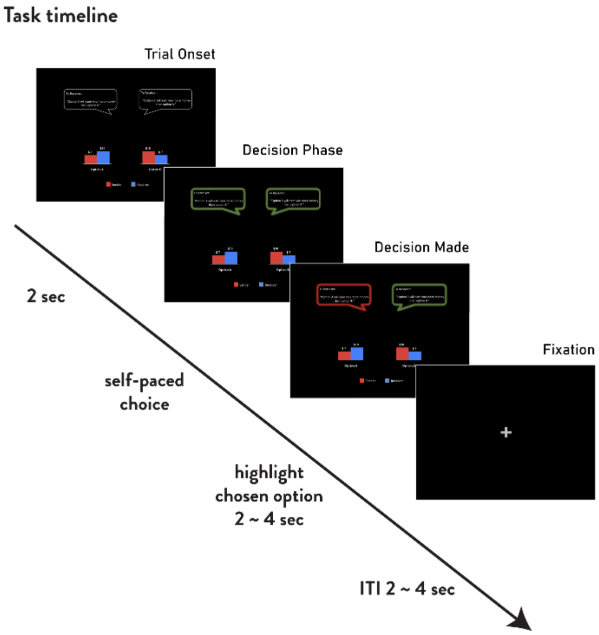


Figure S3. Trial Onset: At the beginning of each trial, all relevant task information was shown on the screen for 2 seconds before the decision phase began. The display included information regarding the four payoffs and the two potential messages that the participant could send. Decision Phase: when the message boxes were highlighted in green, participants were allowed to make their decisions in a self-paced manner. Decision Made: when participants made their decision, their chosen message was highlighted in red to confirm their chosen option. This screen remained present for a period from 2 to 4 seconds, with duration uniformly and randomly distributed. Fixation: a fixation screen appeared during a brief inter-trial interval of 2 to 4 seconds before the next trial.

Option A		Option B	
Self	Other	Self	Other
25	16	10	19
10	10	5	15
15	25	25	15
20	20	5	35
10	25	25	10
10	21	30	19
10	10	15	5
10	25	30	15
19	10	16	25
14	10	15	11
12	8	6	11
15	10	20	9.99
5.99	10	6	5
15	25	16	5
5	30	30	5
19	30	21	10
20	5	15	30
21	10	20	30
35	5	20	20
10	5.99	5	6
8	5	7	10
8	5	5	20
15	25	14.99	30
11	5	10	15
25	15	30	14.99
15	20	10	5
9.99	20	10	15
10	7	5	8
10	21	30	20
5	30	25	28
25	15	11	22
15	25	22	11
19	35	22	5
35	19	5	22
10	21	11	1
25	15	5	16
12	8	11	6
20	10	30	15
20	5	5	8
1	11	21	10
30	15	5	20
5	6	8	10
25	10	15	30
15	10	5	11
28	25	30	5
20	20	10	15

Table S1. Monetary allocations on experimental trials: each subject was presented with the following 46 payoffs twice, once in the deception task, and once in the control task. In 40 out of 46 payoffs, one of the two options was more beneficial to the participant (sender) while the other option was more beneficial to the counterpart (recipient). In 6 out of 46 payoffs, one of the options was beneficial to both sender and recipient. These trials were included as a quality check, as the participants should choose the mutually beneficial option.

Name	Peak MNI coordinates x	 y	 z	Size # voxels (4mm)	Avg. Deception Prediction Mean T-stat (p)	Avg. Generalization T-Ratio (|Preferred T|/ Deception T)	Voxels remaining after correction
L Occipital Cortex	*-18*	*-75*	*6*	*358*	*3.2 (.003)*	*36%*	*229 (64%)*
L Sup. Frontal	*-10*	*28*	*58*	*235*	*3.4 (.002)*	*45%*	*142 (60%)*
R Cuneus	*9*	*-71*	*30*	*174*	*3.2 (.003)*	*29%*	*98 (56%)*
L Superior ACC	*-14*	*40*	*30*	*65*	*3.2 (.003)*	*50%*	*50 (77%)*
L dorsolateral PFC	*-30*	*44*	*34*	*60*	*3.3 (.002)*	*20%*	*48 (80%)*
R Sup. Occipital	*25*	*-87*	*22*	*36*	*3.2 (.003)*	*38%*	*19 (53%)*
L Cuneus	*-14*	*-67*	*30*	*25*	*3.6 (.001)*	*23%*	*16 (64%)*
R Mid. Temporal	*65*	*-11*	*-5*	*25*	*4.2 (.0002)*	*41%*	*18 (72%)*
L Precuneus	*-2*	*-51*	*38*	*24*	*3.2 (.003)*	*48%*	*13 (54%)*
L NAcc	*-6*	*4*	*-9*	*4*	*4.2 (.0002)*	*5%*	*4 (100%)*

Table S2. Peak coordinates for searchlight analysis in Figure 4, sorted by cluster size (descending). Peaks for the deception predictive regions (**Figure 4A**) were identified by searching for local maxima reflecting predictive performance within a 4-voxel radius. The clusters were identified by gradually growing the peaks until all voxels were assigned to one of the 10 identified peaks. The average predictive power for deception within the clusters is reported (corresponding to **Fig 4A**) as well as the average t-statistic ratio, which measures the relative strength of over-generalization compared to prediction performance for deception (corresponding to **Fig 4B**). Finally, the number of voxels that, after dual-goal tuning correction, can still significantly predict deception at uncorrected p < .05 level is reported (corresponding to **Fig 4C**). The p-values are provided as an intuitive conversion of the t-statistic, but because the t-statistics were averaged in the clusters, the p-values do not represent any specific null hypothesis.

Supporting Information References

1. O. Esteban, *et al.*, fMRIPrep: a robust preprocessing pipeline for functional MRI. *Nat Methods* (2019). https://doi.org/10.1038/s41592-018-0235-4.

2. S. Lee, E. T. Bradlow, J. W. Kable, Fast construction of interpretable whole-brain decoders. *Cell Reports Methods* 100227 (2022).

3. S. Lee, *et al.*, A neural signature of the vividness of prospective thought is modulated by temporal proximity during intertemporal decision making. *Proceedings of the National Academy of Sciences* **119**, e2214072119 (2022).

4. M. H. A. Hendriks, N. Daniels, F. Pegado, H. P. Op de Beeck, The effect of spatial smoothing on representational similarity in a simple motor paradigm. *Front Neurol* **8**, 222 (2017).

5. L. Zhu, *et al.*, Damage to dorsolateral prefrontal cortex affects tradeoffs between honesty and self-interest. *Nat Neurosci* **17**, 1319–1321 (2014).

6. U. Gneezy, Deception: The role of consequences. *American Economic Review* **95**, 384–394 (2005).