

Semi-orthogonal subspaces for value mediate a binding and generalization trade-off

Received: 31 August 2023

Accepted: 9 August 2024

Published online: 17 September 2024

 Check for updates

W. Jeffrey Johnston^{1,5}✉, Justin M. Fine^{2,5}, Seng Bum Michael Yoo³,
R. Becket Ebitz⁴ & Benjamin Y. Hayden²

When choosing between options, we must associate their values with the actions needed to select them. We hypothesize that the brain solves this binding problem through neural population subspaces. Here, in macaques performing a choice task, we show that neural populations in five reward-sensitive regions encode the values of offers presented on the left and right in distinct subspaces. This encoding is sufficient to bind offer values to their locations while preserving abstract value information. After offer presentation, all areas encode the value of the first and second offers in orthogonal subspaces; this orthogonalization also affords binding. Our binding-by-subspace hypothesis makes two new predictions confirmed by the data. First, behavioral errors should correlate with spatial, but not temporal, neural misbinding. Second, behavioral errors should increase when offers have low or high values, compared to medium values, even when controlling for value difference. Together, these results support the idea that the brain uses semi-orthogonal subspaces to bind features.

During choice behavior, we must link the anticipated values of each option with the action used to select that option^{1–3}. The unresolved problem of action–value binding is fundamental to neuroeconomics^{3–10}. Moreover, the neural mechanisms underlying value–action binding could apply to other binding problems, such as perceptual binding^{11,12}.

One possible implementation of binding is found in the coding power of neural populations^{13–17}. This power is partly attributable to ‘nonlinear mixed selectivity’^{18–20}. Nonlinear mixed selectivity is conjunctive, in that a neuron’s response depends on the interaction of multiple features (for example, a neuron that responds only to red squares rather than to the color red or the shape square in isolation). Nonlinear mixed selectivity supports flexible behavior in complex cognitive tasks²⁰ and typically provides more reliable representations than linear selectivity^{21–24}. Further, theoretical work shows that conjunctive responses facilitate decoding of multiple stimuli²⁵. Building

on this work, we hypothesize that neural systems bind value to action by encoding the value of options in subspaces of the full population space that are neither perfectly orthogonal nor perfectly parallel to each other. These semi-orthogonal representations are produced by neurons with mixtures of linear and nonlinear mixed selectivity^{19,26}.

Semi-orthogonal representations blend the features of parallel and orthogonal ones. Parallel representations of offers at different positions are low dimensional and factorized. They can unambiguously represent the value and position of a single offer, but, for multiple offers, cannot convey which offer value was at which position. On the other hand, fully orthogonal subspaces preserve this value–position binding, but they have two drawbacks. First, each position requires its own subspace, which becomes impractical when many distinct positions must be represented (see ‘The dimensionality required by different solutions to the binding problem’ in Supplementary Information).

¹Center for Theoretical Neuroscience and Mortimer B. Zuckerman Mind, Brain, and Behavior Institute, Columbia University, New York, NY, USA.

²Department of Neurosurgery, Baylor College of Medicine, Houston, TX, USA. ³Department of Biomedical Engineering, Sunkyunkwan University, and Center for Neuroscience Imaging Research, Institute of Basic Sciences, Suwon, Republic of Korea. ⁴Department of Neuroscience, Université de Montréal, Montreal, Quebec, Canada. ⁵These authors contributed equally: W. Jeffrey Johnston, Justin M. Fine. ✉e-mail: wjeffreyjohnston@gmail.com

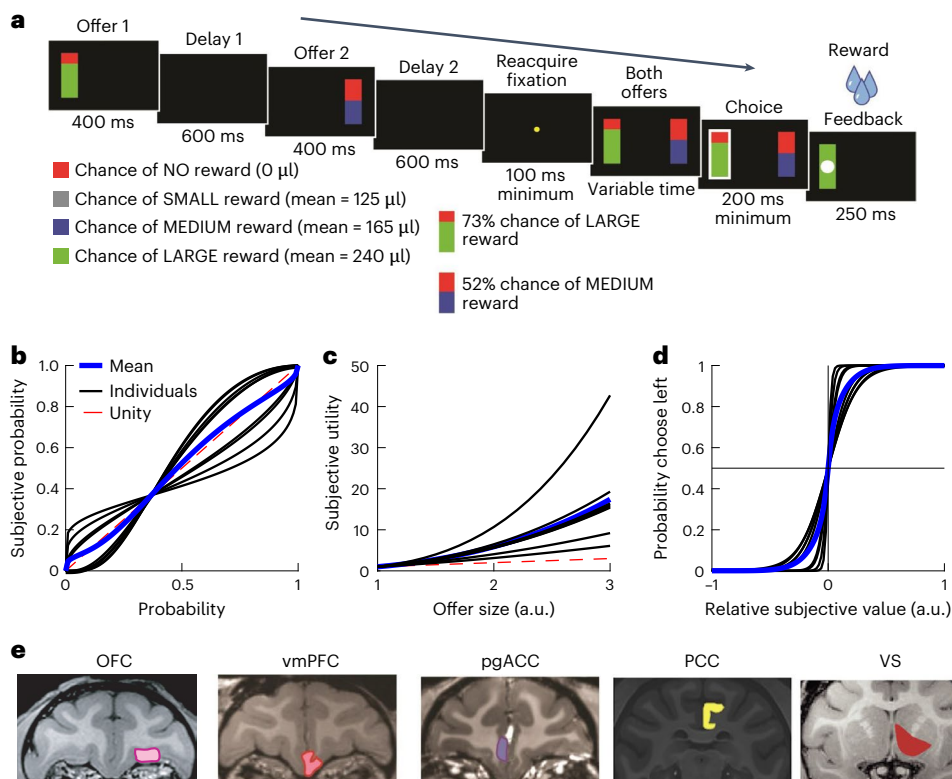


Fig. 1 | Task outline, behavioral model and brain areas. a, Risky-choice task. In the first 400 ms, subjects see the first offer as a bar presented on either the left or the right side. The offer is followed by a 600-ms delay, a 400-ms offer 2, another delay and then choice. **b–d**, Results from the fully subjective behavioral model. **b**, The nonlinear, subjective probability functions fit by the model for each subject (black lines) and their average (blue), compared to the null linear probability

(red unity line). **c**, The same as **b** but for subjective utility. **d**, The probability that each subject chooses the left option given a particular difference in model-fit subjective value, composed of the subjective probability and utility functions in **b** and **c**. For more information on the models, see ‘Behavioral models’ in the Supplementary Information. **e**, Coronal slices from magnetic resonance imaging showing the six different core reward regions that were analyzed. a.u., arbitrary units.

Second, orthogonal representations sacrifice abstraction^{27,28}. That is, a decoder for value trained at one position would not generalize to any other position, making transfer learning nearly impossible^{29,30}. Semi-orthogonal codes occupy a middle ground that avoids both drawbacks^{20,27,31}.

Here, we examined the representational geometry of neural activity in a two-option choice task in five core reward areas: orbitofrontal cortex (OFC), pregenual anterior cingulate cortex (pgACC), posterior cingulate cortex (PCC), ventromedial prefrontal cortex (vmPFC) and ventral striatum (VS). We find that all regions encode the values of left and right offers in distinct, semi-orthogonal subspaces. We then develop a mathematical theory that links subspace orthogonality, binding and abstraction of value to measurable aspects of within-region representational geometry. Using this theory, we show that OFC, pgACC and vmPFC all have representational geometry that is consistent with a near-even trade-off between misbinding errors and generalization errors, while VS minimizes generalization errors and PCC minimizes misbinding.

We then show that, when offers are held in memory, remembered values of first and second offers are encoded in orthogonal subspaces, indicating that the brain uses subspaces to link offer values to the time of their presentation. The geometry of this combined representation makes specific predictions about behavior, allowing us to more rigorously test the binding hypothesis. First, we show that error trials are associated with population activity fluctuations that resemble a spatial, but not temporal, misbinding error—consistent with the fact that the subjects eventually make a spatial choice. Second, we find that stimulus conditions that are less distinct in representational geometry (high-magnitude and low-magnitude offer pairs, as opposed to medium-magnitude pairs) are associated with reduced behavioral accuracy relative to trials that

are more distinct in the geometry. Together, these results support the hypothesis that the brain makes use of subspace orthogonalization to solve the value–action binding problem.

Results

We analyzed data collected from six rhesus macaques performing a two-option, asynchronous risky gambling task that we have used previously³² (Fig. 1). On each trial, subjects use a saccade to select one of two risky offers. Offers occur in sequence on left and right sides of a monitor (400 ms and 600 ms blank, repeated). Then, the two offers reappear simultaneously and the subject chooses. Offers are defined by probability (0–100%, 1% increments) and stakes (0.240 ml or 0.165 ml of juice). Probabilities and stakes are chosen randomly for each option. Order of presentation (left first versus right first) is randomized by trial.

Subject behavior was consistent with previous reports^{33,34}. All subjects preferred offers with higher expected values and were risk-seeking³⁵. Subjects differed in how they weighted probability and stakes. Confirming and extending our previous work, we found that subjects’ choices are best fit by nonlinear, subjective weightings of probability and stakes (Fig. 1b–d; see ‘Behavioral models’ in Supplementary Information and Supplementary Figs. 1 and 2)³³. Our models predict choices with 84–88% accuracy; expected value predicts choices with 70–78% accuracy; models with a single nonlinear term for probability or stakes (but not both) predicted choice with 78–81% and 82–85% accuracy, respectively (Supplementary Fig. 2). The fully nonlinear model has a posterior probability of 0.99 across subjects, indicating a far better fit than the alternative models (Supplementary Fig. 1). Hereafter, we use ‘value’ to mean the model-derived session-specific subjective value. As in our past work, we found little to no satiation effect across sessions (Extended Data Fig. 1a)^{35,36}.

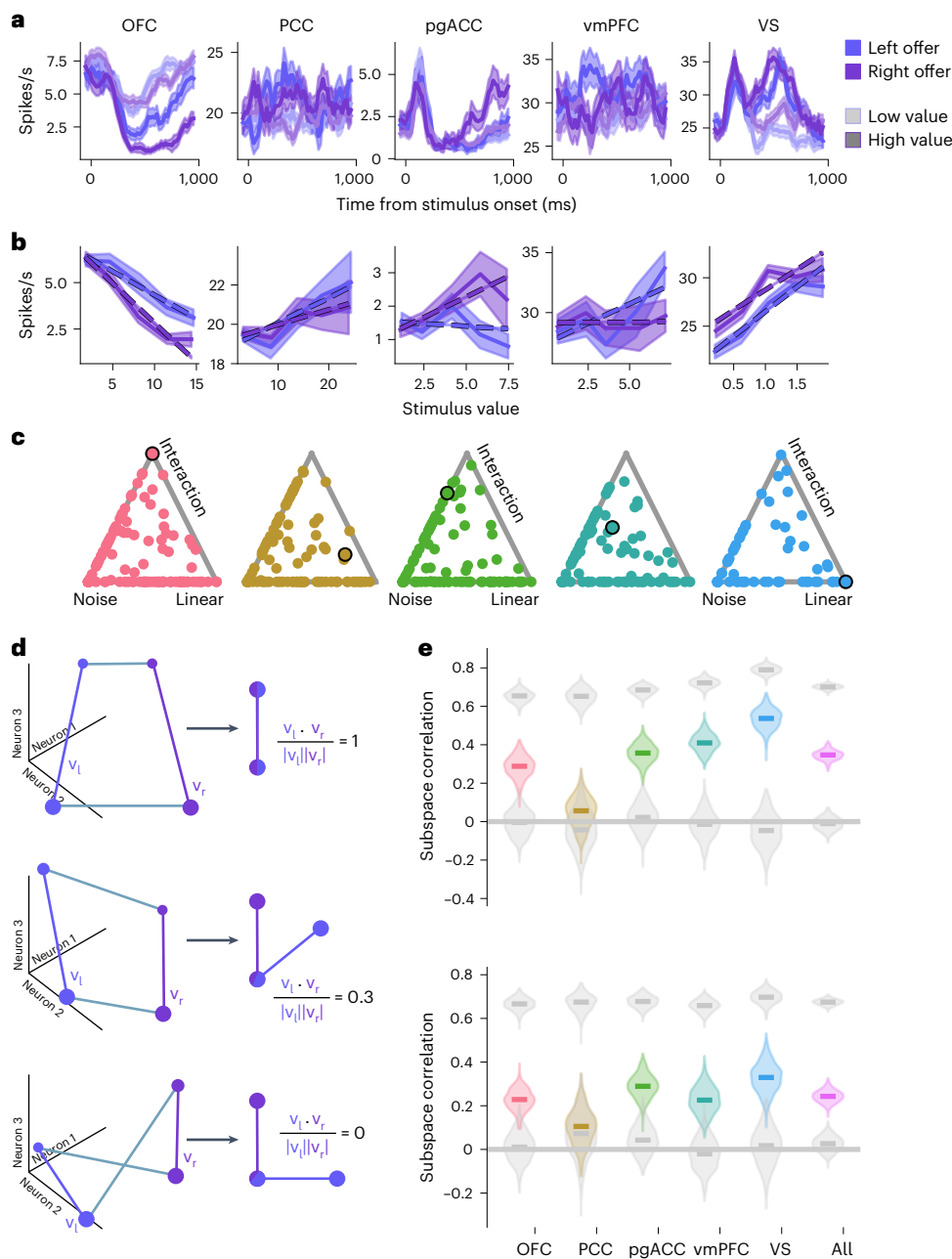


Fig. 2 | Diverse value–response functions produce semi-orthogonal subspaces for value. **a**, The firing rates of example neurons from each region during the offer window, shown for high-value and low-value offers presented on the left or right side (100-ms boxcar filter, shaded area is the s.e.m.). **b**, The value–response function for each neuron in **a**. The value–response function fit by the linear regression model with an interaction term is overlaid (dashed lines; see Extended Data Fig. 2b for the same neurons shown with nonlinear value–response functions). **c**, A simplex showing the weight given to the noise-only, linear and interaction regression models by the Bayesian model stacking analysis (a point in one of the corners indicates that the neuron is best explained by the corresponding model alone). A point in the center of the triangle indicates that the neuron is best explained by an equally weighted combination of all three models. The neurons from **a** and **b** have dark outlines here. Both the linear and interaction categories include models with both linear and spline representation of value. **d**, Schematic of

three different representational geometries that would lead to different subspace correlation results. Top, two perfectly aligned value vectors v_l and v_r in population space (left) would produce a subspace correlation close to 1 (right). Middle, two partially aligned value vectors v_l and v_r would produce a subspace correlation between 0 and 1 (note there is an additional possibility: partially aligned but negatively correlated subspaces; not schematized). Bottom, two unaligned value vectors v_l and v_r would produce a subspace correlation close to 0. **e**, Subspace correlations computed across the full pseudopopulation recorded from each region for the offer presentation window (top) and the delay period (bottom). The rightmost ‘all’ data point is the full recorded pseudopopulation, combined across all regions. The upper gray point is the subspace correlation expected if the left- and right-value subspaces were aligned and corrupted only due to noise; the lower gray point is the subspace correlation obtained from shuffled data. The violin plots show the distribution across the 1,000 bootstrap resamples of each quantity.

Single neurons have nonlinear tuning for value and space

We analyzed the responses of 929 neurons across five brain regions; two subjects per region (Fig. 1; $n = 242$ neurons in OFC, $n = 156$ in vmPFC, $n = 255$ in pgACC, $n = 152$ in PCC and $n = 124$ in VS). Neurons

were recorded with single-contact, high-impedance microelectrodes (vmPFC and VS), or with Plexon u-Probes (pgACC, OFC and PCC). We observed diverse responses from neurons in each region (Fig. 2a), which give rise to diverse value–response functions (Fig. 2b), many of

which are consistent with our hypothesis of a nonlinear interaction between value and space (for example, the example neurons from OFC and pgACC).

To investigate whether these neural responses were best explained by linear or nonlinear interactions between representations of value and space, we fit several linear regression models: (1) a model explaining neural responses in terms of noise (Fig. 2c; 'noise'); (2) a model with separate terms for value and space (Fig. 2c; 'linear'); (3) a model with terms for value, space and a nonlinear interaction between them (Fig. 2c; 'interaction'). For both the linear and interaction models, we fit versions of the model with both linear and spline-based representations for value (see 'Single-neuron linear regression model' in Methods and Extended Data Fig. 2). We compared model fits to the data through a Bayesian model stacking analysis based on approximate leave-one-out cross-validation (see 'Comparison between models with parallel and nonparallel subspaces' in Methods; Fig. 2c). In every brain region, we found substantial proportions of neurons with responses best fit by the nonlinear interaction model (OFC, 12%; PCC, 19%; pgACC, 10%; vmPFC, 12%; VS, 13%; all these proportions are significantly greater than the base rate of 5%, $P < 0.01$, binomial test).

In these neurons, space does not merely shift the value tuning curves of individual neurons, it changes them qualitatively (see Fig. 2b for linear tuning curves and Extended Data Fig. 2b for nonlinear tuning curves). These neurons are not predicted by past work and are inconsistent with previous gain-based proposals (including our own^{37,38}). We also found substantial fractions of neurons whose responses were best explained by the model with only linear selectivity for value and positions (OFC, 27%; PCC, 19%; pgACC, 20%; vmPFC, 25%; VS, 23%; all these proportions were significantly greater than the base rate of 5% with $P < 0.01$, binomial test). The remainder of neurons were best explained by the noise model (OFC, 61%; PCC, 62%; pgACC, 69%; vmPFC, 64%; VS, 64%; all greater than 5%, $P < 0.01$, binomial test); notably, this fraction is similar across brain regions.

Offers are represented in partially overlapping subspaces

What do these single-neuron results mean at the population level? To quantify the degree of overlap between the left- and right-value subspaces, we used the coefficients from the regression model (see above), which define a value-encoding subspace for the left (\mathbf{v}_l) and right (\mathbf{v}_r) sides (Fig. 2d). See Extended Data Fig. 2 for a replication of these results when the subspaces are not constrained to be vectors.

A correlation between left and right vectors that is close to the noise ceiling (gray in Fig. 2e) indicates linear interactions (Fig. 2d, top). The resulting coding scheme cannot implement subspace binding. Subspace correlations less than the noise ceiling are consistent with our hypothesis of subspace binding. A value greater than zero (Fig. 2d, middle) indicates a mixture of both linearly and nonlinearly interacting value and position representations. The intermediate level of subspace correlation (that is, between 1 and 0) supports binding and allows for the generalization of the value code across offer positions and epochs^{27,28,30}. Our analysis approach could also detect zero (Fig. 2d, bottom) or negative (not schematized) subspace correlation. When we use the three terms orthogonal, semi-orthogonal or parallel in the rest of the paper, we are referring to relationships between the offer value subspaces (for example, Fig. 2d, the two purple lines), not the relationship between offer value and offer position subspaces (for example, Fig. 2d, blue and purple lines).

In all brain regions, subspaces for left and right offers are less correlated than the noise ceiling (400 ms offer presentation window; $P < 0.001$ in all cases, Fig. 2e, top; and see 'Computing population encoding subspaces for space and value' in Methods for reliability control). Subspace correlations in every region except PCC are also greater than a shuffle-based floor ($P < 0.001$, Fig. 2e, top; and see 'Computing population encoding subspaces for space and value' in Methods).

We found similar results when repeating the above comparison for the delay period (all $P < 0.05$, Fig. 2e, bottom; for the full time course, see Extended Data Fig. 3e). These results do not reflect heterogeneity between the representations of the first and second offer (Extended Data Fig. 3a). The result replicates in individual subjects, with one exception. The two subjects recorded from pgACC have different subspace correlations: one subject has semi-orthogonal representations of value, while the other has orthogonal representations of value (Extended Data Fig. 3c). These results cannot be explained by changes in the engagement within the course of a session; subjects' likelihood of choosing the higher-value offer remains constant (Extended Data Fig. 1a) and our results do not depend on sessions with larger deviations (Extended Data Fig. 1b).

Semi-orthogonal subspaces provide binding and generalization

These semi-orthogonal representations have been observed in many other studies^{27,31}. Why are these representations favored? First, we showed that, in the noiseless case, a linear decoder can recover the two value–position pairs from a linear, additive representation constructed using any subspace correlation value less than 1 (and greater than -1). This follows from the fact that so long as the value representations are not perfectly parallel (or antiparallel), the value subspaces (for two offers) describe a two-dimensional space. In this two-dimensional space, there are decoding vectors that are correlated with one offer value subspace but orthogonal to the other. Mathematically, the linear encoding matrix for the two stimuli will have an inverse so long as the vectors that encode the two stimuli are not perfectly parallel (see 'Subspace binding for continuous variables' in Supplementary Information).

Next, we included noise and showed that the total error in the decoded offer values strictly decreased as the absolute value of subspace correlation decreased. Consequently, more orthogonal subspaces will lead to more reliable decoding; a subspace correlation of zero will be most reliable (Fig. 3a). Positive subspace correlation leads to a negative correlation between the errors made by a decoder reconstructing the two offer values. Following previous work^{27,30}, we posit that the nonzero subspace correlations we observe provide an abstract representation of value, which facilitates generalization across contexts and rapid learning.

To investigate generalization, we binarized the continuous offer value variable into two categories: high and low (see 'Binarizing value' in Methods for details). Then, we developed a theory that treats the geometry of these kinds of discrete representations as resulting from two components: (1) a rectangular scaffold, where one axis of the rectangle corresponds to offer value and the other corresponds to offer position (Fig. 3b, yellow lines), and (2) high-dimensional perturbations applied to that scaffold (Fig. 3b, dark purple lines). We refer to the length of the rectangular value axis as the linear distance (Fig. 3b, d_l) and the length of the high-dimensional perturbations as the nonlinear distance (Fig. 3b, d_n). A large linear distance and small nonlinear distance implies a high subspace correlation (Fig. 3c, left); similar distances imply a moderate subspace correlation (Fig. 3c, center); a large nonlinear distance and small linear distance implies a low subspace correlation (Fig. 3c, right).

This formalization predicts different kinds of errors. The overall error rate depends on both linear and nonlinear distances (and thus on subspace correlation; Fig. 3d, left). As subspace correlation decreases, the error rate decreases (Fig. 3a). The overall error rate includes the rate of misbinding errors (Fig. 3d, middle), where the value–position associations are mixed up. Interestingly, our theory shows that the misbinding error rate depends on the nonlinear, but not the linear, distance (see 'Derivation of the binding error rate' in Methods). Thus, while representations with low subspace correlation will have low misbinding error rates, it is possible for representations with high

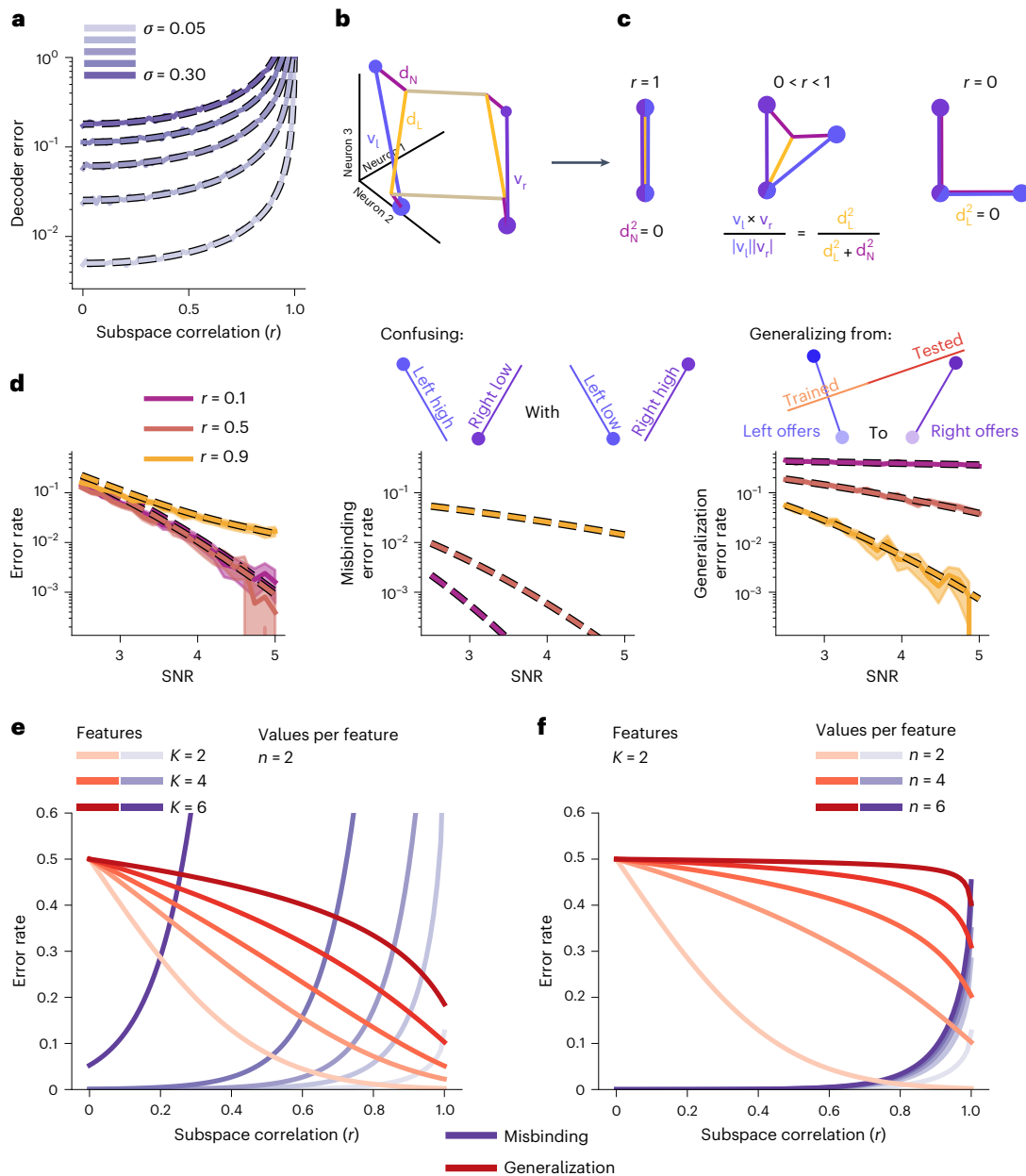


Fig. 3 | Subspace correlation mediates a trade-off between the reliability of binding and generalization. **a**, The overall error rate for continuous representations of multiple offer values at different noise levels and subspace correlations. The dashed line is our analytic theory. **b**, Schematic of the geometric decomposition into linear and nonlinear distances in the discrete case. **c**, The relative magnitude of linear and nonlinear distance determines the subspace correlation used earlier. **d**, The overall error rate (left), misbinding error rate (middle) and generalization error rate (right) for a discrete code in

conditions analogous to the experiment (with two features, $K = 2$, that each take on two values, $n = 2$). The dashed line shows the analytic theory, compared to simulations in the left and right plots. Schematics of misbinding errors (middle) and generalization errors (right) are shown on top of the respective plots. **e**, The dependence of the misbinding (purple) and generalization (red) error rates on the number of features in the stimulus set. **f**, The same as **e** but changing the number of values that each feature takes on. SNR, signal-to-noise ratio.

subspace correlations to have low misbinding error rates, so long as the nonlinear distance is large relative to the noise.

However, we found that the misbinding error rate is in tension with the ability of a decoder for value learned at one location to generalize to a second location. We refer to the error rate at this second, untrained location as the generalization error rate (Fig. 3d, right)²⁷. Our theory shows that the generalization error rate decreases with the linear distance but increases with the nonlinear distance (see ‘Derivation of the generalization error rate’ in Methods). Thus, increasing the nonlinear distance to reduce misbinding errors will increase generalization errors. Both error rates also depend on the properties of the stimulus

set, such as the number of features (Fig. 3e) and number of discrete values (Fig. 3f) that each feature can take on. While misbinding and generalization error rates are always in tension, in many cases low rates of both kinds of error can be achieved by a range of subspace correlations. However, this range becomes smaller as either the number of features (Fig. 3e) or number of values for each feature (Fig. 3f) increases. Increasing the number of features rapidly increases the misbinding error rate (Fig. 3e, purple lines) while having a more moderate effect on the generalization error rate (Fig. 3e, red lines); increasing the number of values rapidly increases the generalization error rate (Fig. 3f, red lines) while having less effect on the misbinding error rate (Fig. 3f,

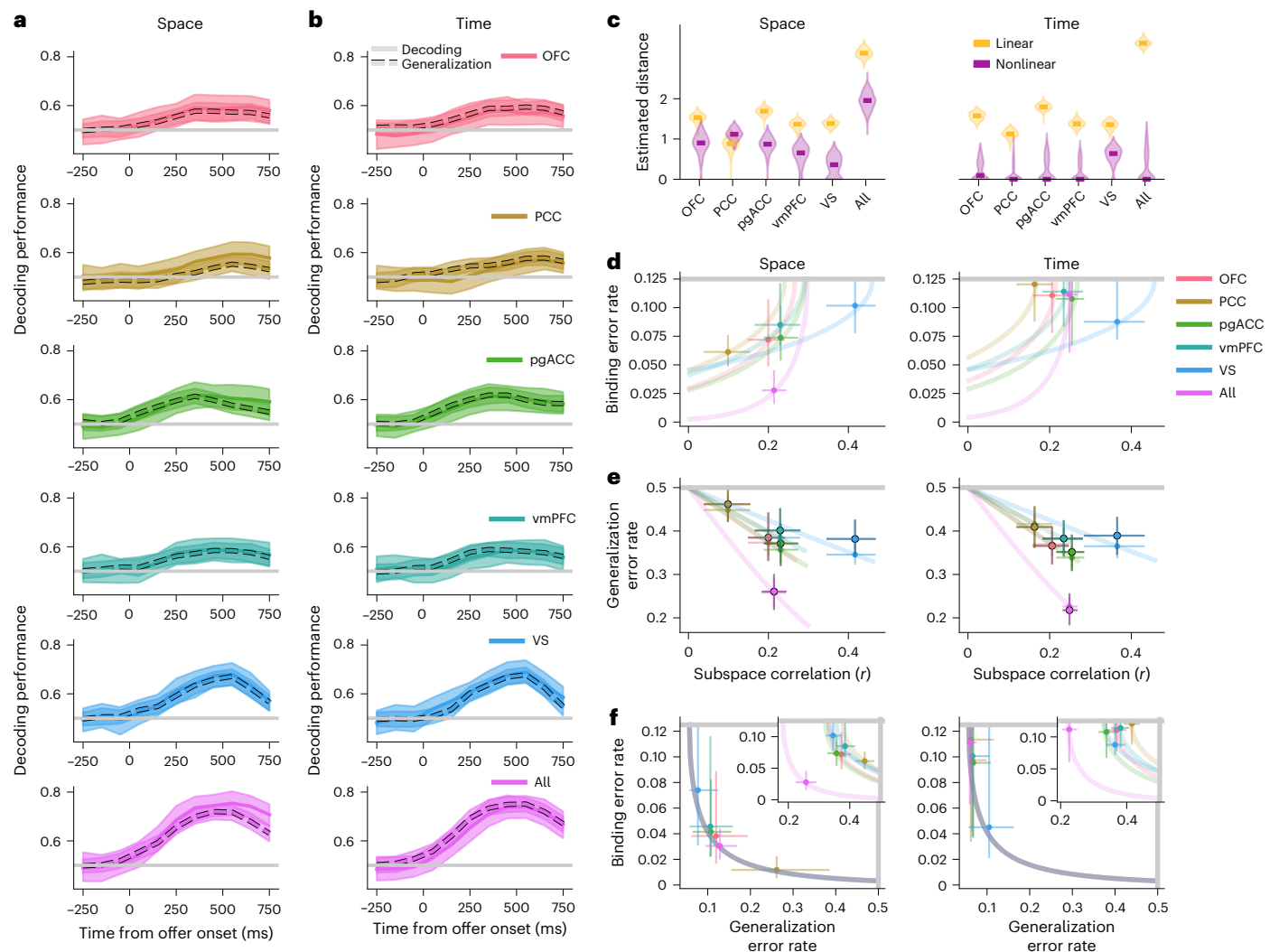


Fig. 4 | The theory predicts misbinding and generalization rates for each region. a, The mean decoding (solid lines) and generalization (dashed lines) performance of decoders trained to decode the binary value of left or right offers. **b**, The same as **a** but for the value of the first or second offer. **c**, The nonlinear and linear distances estimated for the left and right (left) as well as first and second (right) offer value representations within each brain region. The violin plot shows the distribution of bootstrap resamples for the distance estimation procedure. The left and right plots have the same convention for the rest of the figure. The violin plots show the distribution of distance estimates from 200 resamples of the underlying population responses. **d**, The mean predicted binding error rate as a function of subspace correlation for each region, derived from the distance estimates in **c**. The gray line shows the chance

level of binding errors. The extended color line shows the predicted binding rate for codes with the same total power but different trade-offs between linear and nonlinear power. **e**, The mean predicted generalization error rate as a function of subspace correlation for each region, derived from the distance estimates in **c** (for the open circles) and computed empirically with a linear decoder (for the outlined circles). The gray line is the chance level. **f**, Each region shown on the plane defined by the generalization and binding error rates, derived from mean representational power-normalized distance estimates. The inset shows the error rates derived from the unnormalized distance estimates **c**. The gray lines are the chance levels for each of the error types. All error bars and shaded intervals are 95% confidence intervals.

purple lines). Finally, for large numbers of both features and values, it is unlikely that simultaneously low misbinding and generalization error rates cannot be achieved for realistic numbers of neurons and representational power. However, to achieve low misbinding error rates—and in contrast to flexibility-related theories of nonlinear mixed selectivity^{19,20}—a relatively small set of features could nonlinearly interact with all other features, thus providing a kind of nonlinear tag that can be used for binding.

Value representations provide both binding and generalization

Next, we applied the theoretical framework developed in the previous section to the experimental data. First, we decoded the binary value of left and right offers individually, regardless of their time of presentation

(Fig. 4a, solid lines). Second, we decoded the binary value of the first and second offers individually, regardless of their side of presentation (Fig. 4b, solid lines). In both cases, every region yielded decoding accuracies that are above chance starting in the offer presentation window and continuing into the delay period. The decoding performance across different regions was also comparable, although VS had higher decoding performance than vmPFC in the spatial condition during the mid-delay period ($P < 0.05$, Bonferroni corrected, no other region except pgACC survived correction). All regions yielded above-chance cross-condition generalization performance for both the spatial (that is, trained to decode high and low values on left offers, then generalized to right offers; Fig. 4a, dashed lines) and temporal (that is, trained to decode high and low values on offer 1, then generalized to offer 2; Fig. 4b, dashed lines) conditions.

To increase our signal-to-noise ratio and because the representational geometry does not undergo qualitative shifts across the presentation and delay periods (Fig. 4a,b), we combined data across the whole post-presentation window (see ‘Time-pooled decoding’ in Methods). In each of the spatial and temporal conditions, there are four experimental conditions (for the spatial condition: high offer on the left, low offer on the left, high offer on the right, low offer on the right). We estimated the distances between all pairs of the four points using a cross-validated Euclidean distance metric (see ‘Estimating distances’ in Methods), which avoids the positive bias common to many other methods of distance estimation. Then, we decomposed the resulting distance matrix into linear and nonlinear components (see ‘Linear–nonlinear data decomposition’ in Methods and Extended Data Fig. 4). We applied this analysis separately to the left-value and right-value code (averaging over first and second offers) as well as to the first and second offer-value code (averaging over left and right offers). For the space–value representation, we found significant nonlinear distances for all individual brain regions except VS and significant linear distances for all regions (one-sided bootstrap test, $P < 0.05$; Fig. 4c, left). We found significant nonlinear and linear distances when combining neurons across the population (Fig. 4c, left, ‘all’). Through these linear and nonlinear distances, we can develop an independent estimate of subspace correlation (Fig. 4d,e; x axis, and ‘Relating linear and nonlinear distance to subspace correlation’ in Methods).

For the time–value code, we did not find significant nonlinear distances for any region (one-sided bootstrap test, $P > 0.05$), but we did find significant linear distances for every region as well as the population ($P < 0.05$; Fig. 4c, right). We next used these nonlinear distance estimates to predict rate of binding errors. Due to the significant nonlinear distance in every brain region except VS for the space–value code, the predicted misbinding error rate was below chance in every region except VS (Fig. 4d, left). In other words, despite lacking full subspace orthogonality, the population can leverage distinct subspaces to bind offer value to location. As VS has the highest value decoding performance of all recorded regions, we do not believe that this difference between VS and the other regions arises due to a lack of statistical power, even though it has the fewest neurons. Had the problem been one of insufficient power, we would expect VS to have the lowest value decoding performance. The pattern of results is also consistent when the population size for each region is fixed (Extended Data Fig. 5).

However, the time–value representation does not have distinct subspaces for the value of offers presented at different times (Fig. 4d, right), which is inconsistent with binding based on presentation time. That is, first and second offer are represented similarly at their respective time of presentation. However, we show below that the current offer is represented in a distinct subspace from the remembered offer. Thus, time does play a role in binding: offers are bound to their egocentric time of presentation (that is, current or past) rather than their allocentric time of presentation (that is, first or second). As above, distances for pgACC are driven by one of the two subjects, while the results in the other regions are similar across subjects (Extended Data Fig. 6).

Finally, we used these distances to predict the generalization error rate of both the value–space and value–time codes. Due to the significant linear components and moderate (or zero) nonlinear components in every brain region, the predicted generalization error rates are below chance for every brain region (Fig. 4e, open circles). We also computed the empirical generalization error rate of a linear classifier (trained to decode the value of left offers then tested only on right offers). This empirical generalization performance was nearly identical to the predicted performance (Fig. 4e, outlined circles, and see Extended Data Fig. 7 for the standard decoding performance). This agreement indicates that our theory captures the aspects of the representational geometry that are relevant to decoding performance.

To compare the representational geometry of different regions, we normalized the total representational power in each region

(‘Normalizing representational power’ in Methods). This normalization eliminates differences due to different numbers of neurons or different firing rates, while preserving relative linear and nonlinear distances. Next, we computed predicted binding and generalization error rates based on these normalized distances (Fig. 4f; predictions from the original distances are shown in the inset). Interestingly, most regions (and the combined population across all regions) were relatively balanced between the two types of errors. In contrast, PCC specializes for nonlinear representations (predicted binding error rate is less than for the combined population, $P < 0.005$, bootstrap test; predicted generalization error rate is greater than the combined population, $P < 0.005$), while VS specializes for linear representations (predicted binding error rate is greater than for the combined population, $P = 0.01$; predicted generalization error rate is less than the combined population, $P = 0.01$; OFC, vmPFC and pgACC are all no different than the combined population, $P > 0.05$). This supports the idea that VS is a key region for the representation of abstract value³⁹ and PCC may be specialized for spatial representations⁴⁰.

The remembered offer is orthogonal to the current offer

Here, we show that first and second offers are bound to their egocentric time of presentation. A decoder trained to decode the value of offer 1 during its presentation (Fig. 5a, top left, solid line) generalizes to decode the value of offer 2 during its presentation (Fig. 5a, top right, dashed line). Thus, the offer that is currently being presented is encoded in the same subspace whether it is the first or second offer in the trial (Fig. 5b and ‘Time-pooled decoding’ in Methods). However, the remembered value of offer 1 is encoded in a subspace that is orthogonal to this current-offer subspace. A decoder trained to decode the remembered value of offer 1 during the presentation of offer 2 (Fig. 5c, top right, solid line) does not generalize to decode the value of offer 1 while it is being presented (Fig. 5c, top left, dashed line)—and vice versa (Fig. 5c, bottom). This pattern of results holds across the individual regions (although PCC did not reach significance for offer 2; Fig. 5d).

We found that subspace correlations between the past and current offers are indistinguishable from zero in all brain regions ($P < 0.001$ relative to the noise ceiling and $P > 0.05$ relative to zero; Fig. 5e). Together, these results show that the representation of offer 1 (Fig. 5f, left) is in one subspace at the time of presentation, but rotates into a second, orthogonal subspace over the course of the delay period as it presumably enters into working memory (Fig. 5f, middle). Then, when offer 2 is presented, it is represented in the same subspace as the original offer 1 subspace (Fig. 5f, right). This is in keeping with previous work that shows that neural population representations tend to rotate across time^{41,42}. Such rotations can select specific kinds of information for behavioral decisions^{43,44}.

Given the benefits of semi-orthogonality discussed above, it is surprising that we found full orthogonality instead of semi-orthogonality in this context. Further work could explore whether we see consequences of this in the learning and generalization behavior of the animal. This orthogonality could be important to developing a reliable representation of the difference between the two offer values from a particular trial when the animal is getting ready to choose one of the two offers.

The representation of both offers predicts choice behavior

To link the representational geometry of offers with behavior, we estimated the unbiased distance between each pair of (discretized) conditions for neural activity from each region. This yields a representational dissimilarity matrix, an 8×8 symmetric matrix of distances for each region (Fig. 6a, see ‘Estimating distances’ in Methods). If two conditions have a small distance in neural population space, they are more confusable to a decoder.

Three distances are relevant to binding. First, a temporal binding error would occur when trials with a high-value offer on the left followed

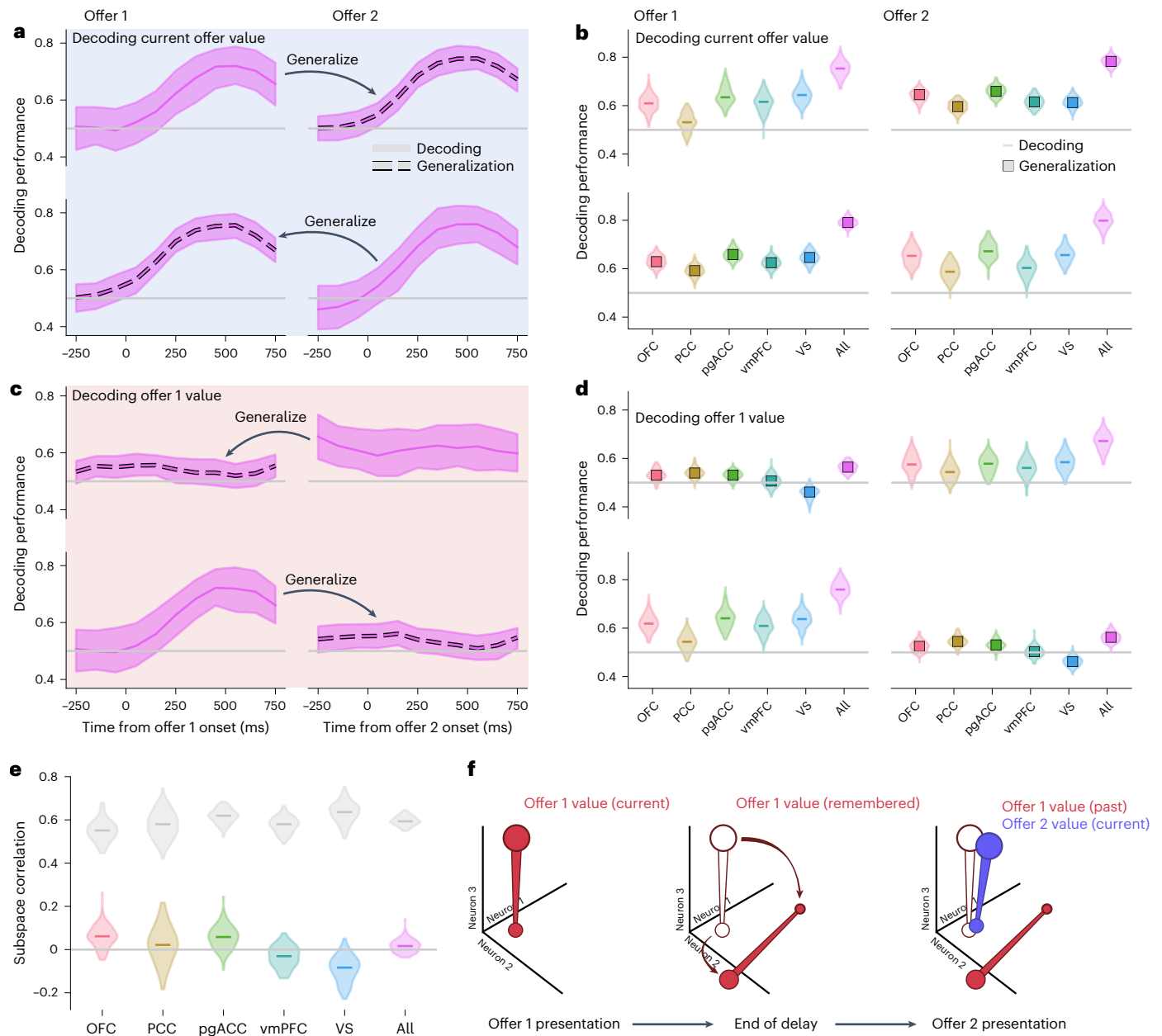


Fig. 5 | The representational geometry of the offer sequence. a, The performance of a classifier trained to decode current-offer value from the combined population. Upper row, the classifier is trained on the presentation of offer 1 (left) and successfully generalizes to offer 2 (right). Lower row, the same as upper, but the decoder is trained on offer 2 (right) and generalizes to offer 1 (left). **b**, The same as **a** but for the individual regions and neural activity is combined across time ('Time-pooled decoding' in Methods). **c**, The performance of a classifier trained to decode the value of offer 1. Upper row, the classifier is trained to decode the value of offer 1 from neural activity recorded during the presentation of offer 2 (right) and generalization performance is tested on neural activity from the presentation of offer 1 (left). Bottom row, the same as

the upper row, but the classifier is trained to decode the value of offer 1 during the presentation of offer 1 (left) and generalization performance is tested on the presentation of offer 2 (right). **d**, The same as **c** but for the individual regions and time-pooled decoding. **e**, The subspace correlation between representations of the current and past offers, computed during the presentation of the second offer and following delay period. **f**, Schematic of what the results from this figure mean for the full geometry of representations during the task. While a single vector is shown here for ease of visualization, this also applies to the rectangular position-value representations shown in Fig. 2. All violin plots show the full distribution across resampled estimates.

by a low-value offer on the right (high left then low right) are confused with low-right then high-left trials (Fig. 6b, left). Second, a spatial binding error would occur when high-left then low-right trials are confused with low-left then high-right trials (Fig. 6b, middle). Third, a combined spatial and temporal binding error occurs when high-left then low-right trials are confused with low-right then high-left trials (Fig. 6b, right). Focusing on these distances revealed that PCC has a distinct geometry from OFC, pgACC, vmPFC and VS. PCC has significant distance between

spatial misbinding conditions, but not temporal misbinding conditions (Fig. 6c; PCC), while OFC and vmPFC have significant distance between temporal but not spatial misbinding conditions (Fig. 6c; OFC and vmPFC). VS, pgACC and the combined population have significant distance between both spatial and temporal misbinding conditions, but significantly larger distance for temporal relative to spatial misbinding conditions (Fig. 6c; VS and 'all'), while PCC has the opposite. Further, we compared the difference between spatial and temporal

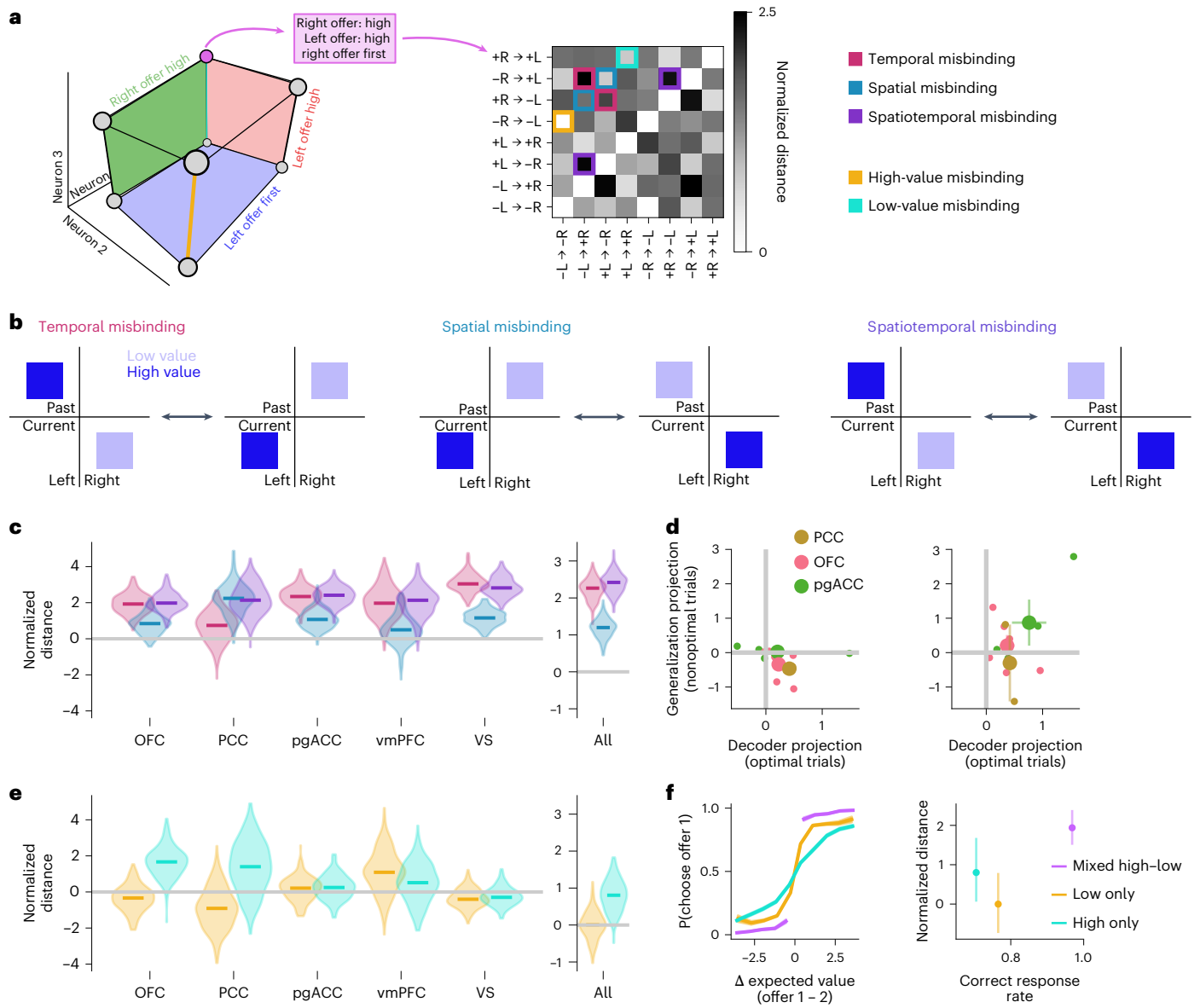


Fig. 6 | The representation of past and current offers predicts key elements of animal behavior. **a**, Left, schematic of the eight experimental conditions, consisting of two offer values and an order of presentation (for example, pink point). Right, the representational dissimilarity matrix estimated for these conditions from the combined population. The colored outlines indicate the distances between condition pairs that are relevant to binding (see **b**). ‘+R’ indicates a high offer on the right side; ‘-L’ indicates a low offer on the left side. The arrow moves from the past to current offer. **b**, The misbinding errors that can occur on this task. Left, two conditions confused in a temporal misbinding error, where the side of the high offer is the same but its time of presentation is different. Middle, two conditions confused in a spatial misbinding error, where the side of the high offer is different but its time of presentation is the same. Right, two conditions confused in a spatiotemporal misbinding error, where the side and presentation of the high offer both differ. **c**, The estimated distances

between the temporal, spatial or spatiotemporal misbinding error conditions. **d**, The mean cross-validated projection along a classifier trained to read out whether the left or right offer was higher (left) and whether the first or second offer was higher (right). The decoder is trained on trials with optimal choices (x axis) and tested on trials with nonoptimal choices (y axis). The small dots represent individual sessions (OFC, $n = 6$; PCC, $n = 2$; pgACC, $n = 4$); the large dot represents the mean across regions, and the error bars are the s.e.m. **e**, The estimated distances between conditions with either both high or both low offers. **f**, Left, the mean behavioral performance across all animals on trials with only high, only low or a mixture of high and low offers. Right, the mean estimated distances from the ‘all’ condition (pooling across $n = 6$ animals) plotted against the mean rate of correct responses in each condition across all sessions. All violin plots show the full distribution across resamples. All error bars are 95% confidence intervals.

misbinding distances across all pairs of regions and found that PCC was significantly different from every other region except OFC ($P < 0.05$, Bonferroni corrected, the OFC–PCC comparison was significant at a nominal level but did not survive the correction). We found that the distance between conditions that would be confused in a spatiotemporal binding error are greater than zero in all regions (one-sided bootstrap test, $P < 0.001$). Because the subjects in our task eventually did make a spatial choice (they chose by shifting their eyes to the left or right),

we predict that trial-to-trial fluctuations along this spatial misbinding dimension may be particularly important for behavior.

To test this prediction, we turned to a trial-by-trial analysis implemented at the experimental session level. We restricted our analysis only to sessions with 10 or more neurons, yielding 12 sessions, 6 from OFC, 2 from PCC and 4 from pgACC—other results from these sessions were consistent with those from all sessions; OFC and pgACC both have semi-orthogonal subspaces for value, while PCC has orthogonal

subspaces (Extended Data Fig. 3d). We trained a decoder to read out whether the value of the left or right offer was higher on time-pooled neural activity following the presentation of the second offer. The decoder was trained only on trials where the animal eventually chooses the higher offer (optimal choices). Then, we computed the average correct margin of held-out trials along the dimension learned by the decoder. First, we computed this distance on held-out trials with optimal choices (Fig. 5f, left, *x* axis). Then, we compute this distance on held-out trials with nonoptimal choices (Fig. 5f, left, *y* axis). The decoders from OFC and PCC had significantly worse margins on trials with nonoptimal choices relative to trials with optimal choices (*t*-test on the difference in average across sessions; OFC, $P = 0.04$; PCC, $P < 0.001$; pgACC, $P = 0.36$; Fig. 5f, left). However, the same relationship did not hold for either region (or for pgACC) when we trained a decoder to read out whether the first or second offer was higher (*t*-test on the difference in average across sessions; OFC, $P = 0.24$; PCC, $P = 0.33$; pgACC, $P = 0.64$; Fig. 5f, right). This result remained stable for other choices of the number of simultaneously recorded neurons (Extended Data Fig. 8). Thus, not only does the representation change on trials with a suboptimal choice, but it also changes specifically toward a spatially (and not temporally) misbound representation. While we only had a handful of sessions for PCC, this result also provides intriguing evidence for a specialized role of PCC in binding offer value to spatial position.

Finally, we considered trials in which both offers are either low or high. In the distance analysis above, trials with a low (high) offer presented on the left followed by a low (high) offer presented on the right were indistinguishable from the representation of a low (high) offer presented on the right followed by a low (high) offer presented on the left (except for high–high offers in OFC and the combined population; Fig. 5g)—in our setting, low (and high) offers are not equivalent and take on many different values. This indicates the trials with both low or both high offers do not strongly bind the specific offer value to time or position. Thus, we predict that the animals may be more likely to make suboptimal choices when both offers are low or high relative to when one offer is high and the other offer is low. So, we reanalyzed the behavior of the animals for low–low and high–high trials compared to low–high or high–low trials and showed that, on average, more suboptimal choices occur for both the low–low and high–high conditions, even when the difference in expected value is the same (Fig. 5h, left). This behavioral finding is consistent with the significantly increased representational distance for high–low and low–high conditions relative to low–low and high–high conditions (bootstrap test, $P < 0.05$ for mixed compared to both high–high and low–low; high–high and low–low are not significantly different from each other; Fig. 5h, right). This pattern of behavior indicates that the animals are sensitive to the absolute magnitude of the presented offers rather than only to their difference. Since the high-value and low-value categories developed here are specific to this task, this suggests that the animals normalize their conception of value to the range of values presented in the current context—as a consequence, we would expect to see this kind of effect in other studies where animals choose between offers with a range of values, even if the absolute values are different.

Discussion

We find that, in a choice task, values of left and right—as well as current and past—offers are represented in distinct subspaces. We develop a theory that predicts both generalization and misbinding errors from the degree of orthogonality between subspaces. We find that OFC, pgACC and vmPFC use codes that balance both error types, while VS minimizes generalization errors and PCC minimizes misbinding errors. Further, neural responses that resemble spatial misbinding are associated with choice errors, as are experimental conditions associated with geometrical ambiguity. These results suggest a new solution to an outstanding problem in neuroeconomics, one grounded in the coding properties of neuronal populations. Moreover, they raise the

possibility that subspace orthogonalization may be a general solution to other important binding problems, such as the perceptual binding problem^{12,13}.

Why are subspaces semi-orthogonal, rather than fully orthogonal? Both orthogonal and semi-orthogonal subspaces allow for binding, but semi-orthogonal subspaces also permit abstraction of value across offer positions²⁷. That is, a decoder learned for value at one position would be able to generalize to a never-before-seen position without retraining³⁰, as well as learn this abstract representation of value from fewer samples than needed for a fully orthogonal code. While previous work has demonstrated a relationship between representational geometry and generalization to novel samples from the same distribution (for example, novel trials, same position)^{21,45}, our work provides a link between this geometry and generalization to samples from a systematically different distribution (for example, novel trials, novel position).

PCC and VS have distinct representational geometries, and our framework predicts that these different geometries would lead to different behavioral consequences if their function is disrupted. For instance, we expect that inactivation of PCC would lead to an increase in the rate of misbinding errors while inactivation of VS would increase the rate of generalization errors or slow transfer learning. Previous theoretical work has shown that semi-orthogonal representations of colors presented at different spatial positions can explain the pattern of errors observed in a common human working memory task²⁵ (but see ref. 46) as well as in a perceptual learning paradigm²⁹—indicating that semi-orthogonal representations may be behaviorally relevant in multiple domains.

While we probed only two positions (and two time points), we predict there will be structure in the degree of orthogonality between representations of stimuli presented at different distances from and times relative to each other. Behavioral work suggests this as well. For instance, when humans are asked to remember an array of color–position pairs, they are more likely to mix up the colors and positions of stimuli that are nearby to each other in space (Fig. 3d, middle; ref. 47). This could arise from less orthogonal representations of those stimuli. Our theory also predicts that estimates of nearby and, therefore, less orthogonal stimuli may have higher variance even when the correct stimulus is reported (Fig. 3a). Further, experimental work has shown that transfer generalization decreases with an increase in the distance between the trained and tested locations⁴⁸; in our framework, this could be due to an increased orthogonality between representations of more distant stimuli. We would expect an analogous pattern of results to hold for stimuli separated in time.

Behavioral investigations of working memory have put forward the idea that distinct stimuli stored in working memory may all be represented in distinct, abstract slots^{49,50}. This slot model has been successful at explaining patterns of errors in working memory recall tasks⁵¹, although there is still debate over whether there is a fixed number of slots or fixed pool of memory resource that can be distributed over any number of items^{52,53}. One advantage of this slots model is that the number of required dimensions (or neurons) scales linearly with the number of slots rather than with the number of positions or presentation times (see ‘The dimensionality required by different solutions to the binding problem’ in the Supplementary Information). However, one notable limitation of the slots model is that it does not provide a clear solution for generalization, because all slots are assumed to be mutually orthogonal. While a linear decoder could read out the same feature from different slots, it would need to be trained on examples from all slots. Further, several of the predictions of the standard slot model are inconsistent with the pattern of orthogonality and semi-orthogonality between distinct offers that we find here (see ‘Predictions for subspace orthogonality given slot-based representations’ in the Supplementary Information and Extended Data Fig. 3). While our data do not rule out all forms of slot-based working memory representations, further work is required to adjudicate between these models.

The result that both low misbinding error rates and above-chance generalization can coexist is important given wide-ranging work about the relative benefits of low-dimensional and high-dimensional neural geometries^{23,54–56}. On one hand, the high-dimensional geometries that arise from nonlinear mixed selectivity have been shown to support flexible decision-making on complex cognitive tasks²⁰. In contrast, relatively low-dimensional abstract, or factorized, representations of different task variables have been shown to support generalization across contexts and rapid learning^{27,28,54,55}. Here, we provide a theory that shows an explicit tension between these benefits—and that shows when the benefits of both forms of representation can be achieved through an intermediate representational geometry, similar to the one observed in both mice and monkeys^{27,31,36,57,58}. Further work can adapt this geometric framework to contexts beyond neuroeconomic decision-making, and use it to generate and test novel predictions about the links between representational geometry and behavior.

These results also build on related work about how the representations of task variables evolve over time^{41,42,59}. We replicate the recent finding that the passage of time is associated with a large-scale rotation in neural dynamics^{41,42}. Thus, in our context, at the time of the second offer presentation, the remembered offer (offer 1) is represented in an orthogonal subspace from the current offer (offer 2), but the current-offer subspace is the same whether that offer comes first or second. Our findings are consistent with the use of rotational dynamics as a scaffold to represent a sequence of stimuli, as proposed by previous work³⁹, although importantly we find evidence for egocentric rather than allocentric temporal representations.

While we have focused on representations of offer value that are tied to specific spatial positions (that is, offers presented on the left and right) and egocentric presentation times (that is, current and past offers), this representation can coexist with representations that depend on or are more closely tied to the animal's eventual choice. In particular, some previous studies have found increased representation of the eventually chosen offer^{32,60} or representations of the difference in value between the two offers, rather than the value of each offer separately³⁹. These representations can exist alongside the representations we focus on here, and would not change the main conclusions of our study.

We and others have noted the encoding of spatial information in single neurons in value-sensitive regions of the brain^{37,61,62}. We previously proposed that these spatial signals modulate the encoding of task variables like value through linear, gain-like changes^{7,37,61}. Our present results indicate, instead, that space alters the tuning profile of value-sensitive neurons. Further, spatial position of the offers and the animal's eventual action were confounded in our task, while recent work shows that the contingency between spatial position and action can also restructure tuning in the frontal cortex⁶³.

The question of how decision-related information such as value is transformed into action is one of the major ones in the field of neuroeconomics^{3,8,64,65}. Typical theories hold that values are represented in an abstract value space that is conceptually and functionally distinct from an action space used to implement choices. As such, there is a question of how value is linked to action. This problem is often reified in neuroanatomy—some regions are assumed to be pure value regions, while other presumably anatomically downstream regions are assumed to have action signals^{1,3,66}. Our results suggest, instead, that the same neural code provides an abstract representation of value and representations of value that are bound to particular spatial positions. Further, we do not find that the neurons supporting these different aspects (abstract as opposed to spatially bound) are divided into distinct subpopulations (Supplementary Fig. 3). This argues against the idea that there is an anatomical distinction between value and action frames^{67,68}. Our findings suggest that this mechanism for binding through semi-orthogonal subspaces is used throughout the cortical reward system. Together, these results highlight the potential value of

functional specialization through population representations, rather than through modular architecture^{15,17}.

Online content

Any methods, additional references, Nature Portfolio reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at <https://doi.org/10.1038/s41593-024-01758-5>.

References

1. Kable, J. W. & Glimcher, P. W. The neurobiology of decision: consensus and controversy. *Neuron* **63**, 733–745 (2009).
2. Samejima, K., Ueda, Y., Doya, K. & Kimura, M. Representation of action-specific reward values in the striatum. *Science* **310**, 1337–1340 (2005).
3. Rangel, A., Camerer, C. & Montague, P. R. A framework for studying the neurobiology of value-based decision making. *Nat. Rev. Neurosci.* **9**, 545–556 (2008).
4. Wunderlich, K., Rangel, A. & O'Doherty, J. P. Neural computations underlying action-based decision making in the human brain. *Proc. Natl Acad. Sci. USA* **106**, 17199–17204 (2009).
5. Cai, X. & Padoa-Schioppa, C. Contributions of orbitofrontal and lateral prefrontal cortices to economic choice and the good-to-action transformation. *Neuron* **81**, 1140–1151 (2014).
6. Hare, T. A., Schultz, W., Camerer, C. F., O'Doherty, J. P. & Rangel, A. Transformation of stimulus value signals into motor commands during simple choice. *Proc. Natl Acad. Sci. USA* **108**, 18120–18125 (2011).
7. Hayden, B. Y. & Moreno-Bote, R. A neuronal theory of sequential economic choice. *Brain Neurosci. Adv.* **2**, 2398212818766675 (2018).
8. Knudsen, E. B. & Wallis, J. D. Taking stock of value in the orbitofrontal cortex. *Nat. Rev. Neurosci.* **23**, 428–438 (2022).
9. Padoa-Schioppa, C. & Assad, J. A. The representation of economic value in the orbitofrontal cortex is invariant for changes of menu. *Nat. Neurosci.* **11**, 95–102 (2008).
10. Yim, M. Y., Cai, X. & Wang, X. J. Transforming choice outcome to action plan in monkey lateral prefrontal cortex: a neural circuit model. *Neuron* **103**, 520–532 (2019).
11. Stoet, G. & Hommel, B. Action planning and the temporal binding of response codes. *J. Exp. Psychol. Hum. Percept. Perform.* **25**, 1625–1640 (1999).
12. Treisman, A. M. & Gelade, G. A feature-integration theory of attention. *Cogn. Psychol.* **12**, 97–136 (1980).
13. Roelfsema, P. R. Solving the binding problem: assemblies form when neurons enhance their firing rate—they don't need to oscillate or synchronize. *Neuron* **111**, 1003–1019 (2023).
14. Chung, S. & Abbott, L. F. Neural population geometry: an approach for understanding biological and artificial neural networks. *Curr. Opin. Neurobiol.* **70**, 137–144 (2021).
15. Ebitz, R. B. & Hayden, B. Y. The population doctrine in cognitive neuroscience. *Neuron* **109**, 3055–3068 (2021).
16. Mante, V., Sussillo, D., Shenoy, K. V. & Newsome, W. T. Context-dependent computation by recurrent dynamics in prefrontal cortex. *Nature* **503**, 78–84 (2013).
17. Saxena, S. & Cunningham, J. P. Towards the neural population doctrine. *Curr. Opin. Neurobiol.* **55**, 103–111 (2019).
18. Blanchard, T. C., Piantadosi, S. T. & Hayden, B. Y. Robust mixture modeling reveals category-free selectivity in reward region neuronal ensembles. *J. Neurophysiol.* **119**, 1305–1318 (2018).
19. Fusi, S., Miller, E. K. & Rigotti, M. Why neurons mix: high dimensionality for higher cognition. *Curr. Opin. Neurobiol.* **37**, 66–74 (2016).

20. Rigotti, M. et al. The importance of mixed selectivity in complex cognitive tasks. *Nature* **497**, 585–590 (2013).
21. Barak, O., Rigotti, M. & Fusi, S. The sparseness of mixed selectivity neurons controls the generalization–discrimination trade-off. *J. Neurosci.* **33**, 3844–3856 (2013).
22. Babadi, B. & Sompolinsky, H. Sparseness and expansion in sensory representations. *Neuron* **83**, 1213–1226 (2014).
23. Litwin-Kumar, A., Harris, K. D., Axel, R., Sompolinsky, H. & Abbott, L. F. Optimal degrees of synaptic connectivity. *Neuron* **93**, 1153–1164 (2017).
24. Johnston, W. J., Palmer, S. E. & Freedman, D. J. Nonlinear mixed selectivity supports reliable neural computation. *PLoS Comput. Biol.* **16**, e1007544 (2020).
25. Matthey, L., Bays, P. M. & Dayan, P. A probabilistic palimpsest model of visual short-term memory. *PLoS Comput. Biol.* **11**, e1004003 (2015).
26. Parthasarathy, A. et al. Mixed selectivity morphs population codes in prefrontal cortex. *Nat. Neurosci.* **20**, 1770–1779 (2017).
27. Bernardi, S. et al. The geometry of abstraction in the hippocampus and prefrontal cortex. *Cell.* **183**, 954–967 (2020).
28. Johnston, W. J. & Fusi, S. Abstract representations emerge naturally in neural networks trained to perform multiple tasks. *Nat. Commun.* **14**, 1040 (2023).
29. Doshier, B. A., Jeter, P., Liu, J. & Lu, Z. L. An integrated reweighting theory of perceptual learning. *Proc. Natl Acad. Sci. USA* **110**, 13678–13683 (2013).
30. Doshier, B. & Lu, Z. L. Visual perceptual learning and models. *Annu. Rev. Vis. Sci.* **3**, 343–363 (2017).
31. Yoo, S. B. M. & Hayden, B. Y. The transition from evaluation to selection involves neural subspace reorganization in core reward regions. *Neuron* **105**, 712–724 (2020).
32. Strait, C. E., Blanchard, T. C. & Hayden, B. Y. Reward value comparison via mutual inhibition in ventromedial prefrontal cortex. *Neuron* **82**, 1357–1366 (2014).
33. Farashahi, S., Azab, H., Hayden, B. & Soltani, A. On the flexibility of basic risk attitudes in monkeys. *J. Neurosci.* **38**, 4383–4398 (2018).
34. Farashahi, S., Donahue, C. H., Hayden, B. Y., Lee, D. & Soltani, A. Flexible combination of reward information across primates. *Nat. Hum. Behav.* **3**, 1215–1224 (2019).
35. Heilbronner, S. & Hayden, B. Contextual factors explain risk-seeking preferences in rhesus monkeys. *Front. Neurosci.* <https://doi.org/10.3389/fnins.2013.00007> (2013).
36. Fine, J. M. et al. Abstract value encoding in neural populations but not single neurons. *J. Neurosci.* **43**, 4650–4663 (2023).
37. Strait, C. E. et al. Neuronal selectivity for spatial positions of offers and choices in five reward regions. *J. Neurophysiol.* **115**, 1098–1111 (2016).
38. Hayden, B. Y. & Platt, M. L. Neurons in anterior cingulate cortex multiplex information about reward and action. *J. Neurosci.* **30**, 3339–3346 (2010).
39. Strait, C. E., Sleezer, B. J. & Hayden, B. Y. Signatures of value comparison in ventral striatum neurons. *PLoS Biol.* **13**, e1002173 (2015).
40. Dean, H. L. & Platt, M. L. Allocentric spatial referencing of neuronal activity in macaque posterior cingulate cortex. *J. Neurosci.* **26**, 1117–1127 (2006).
41. Libby, A. & Buschman, T. J. Rotational dynamics reduce interference between sensory and memory representations. *Nat. Neurosci.* **24**, 715–726 (2021).
42. Pu, S., Dang, W., Qi, X. L. & Constantinidis, C. Prefrontal neuronal dynamics in the absence of task execution. *Nat. Commun.* <https://doi.org/10.1010/2022.09.16.508324> (2022).
43. Panichello, M. F. & Buschman, T. J. Shared mechanisms underlie the control of working memory and attention. *Nature* **592**, 601–605 (2021).
44. Piwek, E. P., Stokes, M. G. & Summerfield, C. A recurrent neural network model of prefrontal brain activity during a working memory task. *PLoS Comput. Biol.* **19**, e1011555 (2023).
45. Sorscher, B., Ganguli, S. & Sompolinsky, H. Neural representational geometry underlies few-shot concept learning. *Proc. Natl Acad. Sci. USA* **119**, e2200800119 (2022).
46. Alleman, M., Panichello, M., Buschman, T. J. & Johnston, W. J. The neural basis of swap errors in working memory. *Proc. Natl Acad. Sci. USA* **121**, e2401032121 (2024).
47. Schneegans, S. & Bays, P. M. Neural architecture for feature binding in visual working memory. *J. Neurosci.* **37**, 3913–3925 (2017).
48. Crist, R. E., Kapadia, M. K., Westheimer, G. & Gilbert, C. D. Perceptual learning of spatial localization: specificity for orientation, position, and context. *J. Neurophysiol.* **78**, 2889–2894 (1997).
49. Luck, S. J. & Vogel, E. K. The capacity of visual working memory for features and conjunctions. *Nature* **390**, 279–281 (1997).
50. Cowan, N. The magical number 4 in short-term memory: a reconsideration of mental storage capacity. *Behav. Brain Sci.* **24**, 87–114 (2001).
51. Zhang, W. & Luck, S. J. Discrete fixed-resolution representations in visual working memory. *Nature* **453**, 233–235 (2008).
52. Ma, W. J., Husain, M. & Bays, P. M. Changing concepts of working memory. *Nat. Neurosci.* **17**, 347–356 (2014).
53. Bays, P. M., Schneegans, S., Ma, W. J. & Brady, T. F. Representation and computation in visual working memory. *Nat. Hum. Behav.* <https://doi.org/10.1038/s41562-024-01871-2> (2024).
54. Cueva, C. J. et al. Low-dimensional dynamics for working memory and time encoding. *Proc. Natl Acad. Sci. USA* **117**, 23021–23032 (2020).
55. Gallego, J. A., Perich, M. G., Miller, L. E. & Solla, S. A. Neural manifolds for the control of movement. *Neuron* **94**, 978–984 (2017).
56. Jazayeri, M. & Ostojic, S. Interpreting neural computations by examining intrinsic and embedding dimensionality of neural activity. *Curr. Opin. Neurobiol.* **70**, 113–120 (2021).
57. Nogueira, R., Rodgers, C. C., Bruno, R. M. & Fusi, S. The geometry of cortical representations of touch in rodents. *Nat. Neurosci.* **26**, 239–250 (2023).
58. Boyle, L., Posani, L., Irfan, S., Siegelbaum, S. A. & Fusi, S. Tuned geometries of hippocampal representations meet the computational demands of social memory. *Neuron* **112**, 1358–1371 (2024).
59. Koay, S. A., Charles, A. S., Thiberge, S. Y., Brody, C. D. & Tank, D. W. Sequential and efficient neural-population coding of complex task information. *Neuron* **110**, 328–349 (2022).
60. Gore, F. et al. Orbitofrontal cortex control of striatum leads economic decision-making. *Nat. Neurosci.* **26**, 1566–1574 (2023).
61. Yoo, S. B. M., Sleezer, B. J. & Hayden, B. Y. Robust encoding of spatial information in orbitofrontal cortex and striatum. *J. Cogn. Neurosci.* **30**, 898–913 (2018).
62. Feierstein, C. E., Quirk, M. C., Uchida, N., Sosulski, D. L. & Mainen, Z. F. Representation of spatial goals in rat orbitofrontal cortex. *Neuron* **51**, 495–507 (2006).
63. Jonikaitis, D. & Zhu, S. Action space restructures visual working memory in prefrontal cortex. Preprint at *bioRxiv* <https://doi.org/10.1101/2023.08.13.553135> (2023).
64. Krajbich, I., Armel, C. & Rangel, A. Visual fixations and the computation and comparison of value in simple choice. *Nat. Neurosci.* **13**, 1292–1298 (2010).

65. Padoa-Schioppa, C. & Assad, J. A. Neurons in the orbitofrontal cortex encode economic value. *Nature* **441**, 223–226 (2006).
66. Padoa-Schioppa, C. Neurobiology of economic choice: a good-based model. *Annu. Rev. Neurosci.* **34**, 333–359 (2011).
67. Fine, J. M. & Hayden, B. Y. The whole prefrontal cortex is premotor cortex. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **377**, 20200524 (2022).
68. Hayden, B. Y. & Niv, Y. The case against economic values in the orbitofrontal cortex (or anywhere else in the brain). *Behav. Neurosci.* **135**, 192–201 (2021).

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.

© The Author(s), under exclusive licence to Springer Nature America, Inc. 2024

Methods

Surgical procedures

All procedures were approved by either the University Committee on Animal Resources at the University of Rochester or the Institutional Animal Care and Use Committee at the University of Minnesota. Animal procedures were also designed and conducted in compliance with the Public Health Service's Guide for the Care and Use of Animals. All surgery was performed under anesthesia. Six male rhesus macaques (*Macaca mulatta*) aged from 3 to 6 years served as subjects. A small prosthesis was used to maintain stability. Animals were habituated to laboratory conditions and then trained to perform oculomotor tasks for liquid rewards. We placed a Cilux recording chamber (Crist Instruments) over the area of interest. We verified positioning by magnetic resonance imaging with the aid of a Brainsight system (Rogue Research). Animals received appropriate analgesics and antibiotics after all procedures. Throughout both behavioral and physiological recording sessions, we kept the chamber clean with regular antibiotic washes and sealed them with sterile caps.

Recording sites

We approached our brain regions through standard recording grids (Crist Instruments) guided by a micromanipulator (NAN Instruments). All recording sites were selected based on the boundaries given in the Paxinos atlas⁶⁹. In all cases, we sampled evenly across the regions. Neuronal recordings in OFC were collected from subjects P and S; recordings in vmPFC were collected from subjects B and H; recordings in pgACC were collected from subjects B and V; recordings from PCC were collected from subjects P and S; and recordings in VS were collected from subjects B and C.

Data for vmPFC and VS were collected with single-tip microwire microelectrodes (FHC). Data from the other regions were collected with multicontact ($n = 16$ or 32) u-Probe microelectrodes (Plexon).

We defined OFC 11/13 as lying within the coronal planes situated between 28.65 mm and 42.15 mm rostral to the interaural plane, the horizontal planes situated between 3 mm and 9.5 mm from the brain's ventral surface, and the sagittal planes between 3 mm and 14 mm from the medial wall. The coordinates correspond to both areas 11 and 13 in Paxinos et al.⁶⁹. We used the same criteria in a different dataset⁷⁰. We recorded from OFC using one Plexon u-Probe multicontact microelectrode (16 channels) a total of seven times, yielding a median of 32 neurons per session (range: 9–59) for a total of 242 neurons.

We defined vmPFC 14 as lying within the coronal planes situated between 29 mm and 44 mm rostral to the interaural plane, the horizontal planes situated between 0 mm and 9 mm from the brain's ventral surface, and the sagittal planes between 0 mm and 8 mm from the medial wall. These coordinates correspond to area 14 m in Paxinos et al.⁶⁹. This dataset was used in Strait et al.^{32,37}. We recorded from vmPFC 138 times, yielding a median of 1 neuron per session (range: 1–3) for a total of 156 neurons.

We defined pgACC 32 as lying within the coronal planes situated between 30.90 mm and 40.10 mm rostral to the interaural plane, the horizontal planes situated between 7.30 mm and 15.50 mm from the brain's dorsal surface, and the sagittal planes between 0 mm and 4.5 mm from the medial wall. Our recordings were made from central regions within these zones, which correspond to area 32 in Paxinos et al.⁶⁹. Note that the term 'area 32' is sometimes used more broadly than we use it here. In those studies, it encompasses the dorsal anterior cingulate cortex, which we did not study here and believe should be called 'area 24'⁷¹. We recorded from pgACC a total of 115 times, yielding a median of 1 neuron per session (range: 1–35) for a total of 255 neurons.

We defined PCC 29/31 as lying within the coronal planes situated between 2.88 mm caudal and 15.6 mm rostral to the interaural plane, the horizontal planes situated between 16.5 mm and 22.5 mm from the brain's dorsal surface, and the sagittal planes between 0 and 6 mm from the medial wall. The coordinates correspond to area 29/31 in

Paxinos et al.^{69,72}. We recorded from PCC twice, yielding 63 and 89 neurons in each of the two sessions for a total of 152 neurons.

We defined VS as lying within the coronal planes situated between 20.66 mm and 28.02 mm rostral to the interaural plane, the horizontal planes situated between 0 mm and 8.01 mm from the ventral surface of the striatum, and the sagittal planes between 0 mm and 8.69 mm from the medial wall. Note that our recording sites were targeted toward the nucleus accumbens core region of the VS. This dataset was used in Strait et al.^{37,39}. We recorded from VS a total of 49 times, yielding a median of 2 neurons per session (range: 1–7) for a total of 124 neurons.

We confirmed the recording location before each recording session using our Brainsight system with structural magnetic resonance images taken before the experiment. Neuroimaging was performed at the Rochester Center for Brain Imaging on a Siemens 3T MAGNETOM Trio Tim using 0.5-mm voxels or in the Center for Magnetic Resonance Research at University of Minnesota. We confirmed recording locations by listening for characteristic sounds of white and gray matter during recording, which in all cases matched the loci indicated by the Brainsight system.

Electrophysiological techniques and processing

Either single (FHC) or multicontact (32-channel V-Probe, Plexon) electrodes were lowered using a microdrive (NAN Instruments) until waveforms could be isolated. Up to four electrodes (V-probe and/or FHC) were inserted simultaneously during a single session. Individual action potentials were isolated on a Plexon system (Plexon) or Ripple Neuro. Neurons were selected for study solely on the basis of the quality of isolation; we never preselected based on task-related response properties. All collected neurons for which we managed to obtain at least 300 trials were analyzed; no neurons that surpassed our isolation criteria were excluded from analysis.

Eye-tracking and reward delivery

Eye position was sampled at 1,000 Hz by an infrared eye-monitoring camera system (SR Research). Stimuli were controlled by a computer running MATLAB 2012 (MathWorks) with Psychtoolbox 3.0 and EyeLink Toolbox. Visual stimuli were colored rectangles on a computer monitor placed 57 cm from the animal and centered on its eyes. A standard solenoid valve controlled the duration of juice delivery. Solenoid calibration was performed daily.

Risky-choice task

The task made use of vertical rectangles indicating reward amount and probability. We have shown in a variety of contexts that this method provides reliable communication of abstract concepts such as reward, probability, delay and rule to monkeys^{73–75}. The task presented two offers on each trial. A rectangle 300 pixels tall and 80 pixels wide represented each offer (11.35° of visual angle tall and 4.08° of visual angle wide). Two parameters defined gamble offers—stakes and probability. Each gamble rectangle was divided into two portions, one red and the other gray, blue or green. The size of the color portions signified the probability of winning a small (125 μ l, gray), medium (165 μ l, blue) or large reward (240 μ l, green), respectively. We used a uniform distribution between 0% and 100% for probabilities. The size of the red portion indicated the probability of no reward. Offer types were selected at random with a 43.75% probability of blue (medium magnitude) gamble, a 43.75% probability of green (high magnitude) gambles, and a 12.5% probability of gray options (safe offers). All safe offers were excluded from the analyses described here, although we confirmed that the results are the same if these trials are included. Previous training history for these subjects included several saccade-based laboratory tasks, including a foraging task⁷⁵, two stochastic choice tasks^{76,77}, a foraging task⁷⁰ and a card sorting task⁷⁸.

On each trial, one offer appeared on the left side of the screen and the other appeared on the right. We randomized the sides of the first

and second offer. Both offers appeared for 400 ms and were followed by a 600-ms blank period. After the offers were presented separately, a central fixation spot appeared, and the monkey fixated on it for 100 ms. Next, both offers appeared simultaneously and the animal indicated its choice by shifting gaze to its preferred offer and maintaining fixation on it for 200 ms. Failure to maintain gaze for 200 ms did not lead to the end of the trial but instead returned the monkey to a choice state; thus, monkeys were free to change their mind if they did so within 200 ms (although in our observations, they seldom did so). Following a successful 200-ms fixation, the gamble was resolved and the reward was delivered. We defined trials that took >7 s as inattentive trials and we did not include them in the analyses (this removed ~1% of trials). Outcomes that yielded rewards were accompanied by a visual cue: a white circle in the center of the chosen offer. All trials were followed by an 800-ms intertrial interval with a blank screen.

Behavioral control for neural data indicates decisions from online valuation comparison

The aim in this task design is to examine the process of valuation and decision-making in a sequential setting. In our design, the subjects are presented with the offers again to allow them the chance to reify their choice. We have trained monkeys to perform a version of the task that is identical except that the simultaneous presentation does not occur, and the animal must choose an offer side from memory. In behavioral data from this modified task (seven sessions for four subjects, previously reported in ref. 31), we found that the monkeys chose the higher-value offer with a similar probability (main task, 82%; modified task, 80%). While this observation does not guarantee that the monkeys use the sequential presentations in the task here to plan a choice, it provides evidence that the decision process is similar when the animal is only shown a sequential presentation. Further, due to the animal's ability to freely saccade during the sequential presentation, we know that they attend to both stimuli (reported on in detail in ref. 79), consistent with finding representations of both value and position during this presentation window.

Behavioral analysis, model estimation and subjective value

The decision variables underlying subject choice could arise from several possible estimates over probability, stakes and the estimated value. Our previous analysis and modeling of the present behavioral data indicate that the monkeys in this task make choices in line with a subjective valuation estimate of offers that reflects subjective attitudes toward the offer size (stakes) that indicate risk-seeking, and a warped probability estimation fit well by a Prelec warping function. We remodeled the decisions here to derive subjective value estimates for using in estimation of neural encoding subspaces (see below) and establishing the connection of these subspaces to monkey choices. Here, we considered models where the monkey's subjective value for an offer follows the probability times the stakes.

The model space we considered included four models. The models included those where the stakes were either assumed to be observed objectively or weighted as a power law, and the probability was either observed objectively or transformed with a Prelec function. For all models, choice probabilities were assumed to be generated by a softmax decision function over the relative subjective value. For example, during model fitting of a model with subjective utility (weighted stakes) and weighted probability, three model terms were fit—the terms for utility and probability, and the softmax temperature term. All models were fitted using the variational Bayesian toolbox, and model comparisons were performed using Bayesian model selection over the model free energies⁸⁰.

Full set of trial conditions for binary value

The full simultaneous representation of the past and current offer can be described by eight distinct conditions, when we binarize offer

value as before (Methods). The values of the first and second offers are chosen independently—so, there are four possible combinations of the two offer values: both low, both high, and two combinations where one is high and the other is low (that is, the left offer is high or the right offer is high). Then there are two possible orders: either the first offer is presented on the left (which means the second offer is presented on the right) or the first offer is presented on the right (and the second offer is presented on the left). This gives us eight conditions.

Neural processing, data selection and statistical analysis

We calculated firing rates in 20-ms bins, but we analyzed them in longer (400-ms and 500-ms) epochs. For estimating regression coefficients (see below), firing rates were all z-scored for each neuron. In the regression, decoding and distance analyses, we only included risky trials, although we provide a figure showing that our main results are replicated when these trials are included (Extended Data Fig. 9). We made the decision to exclude them in the main text in part due to recent work showing that 'safe' options in this kind of task are represented in a distinct way from risky trials³⁶. There is one exception to this, however: we included safe trials when estimating the distances between all eight experimental conditions (Fig. 6). In that regime, we found it was necessary to include the safe trials to have enough trials per condition to estimate the distances.

Single-neuron linear regression model. Individual neuron selectivity was estimated using a linear regression model. Model coefficients were estimated using the z-scored, time-averaged firing rate in four different task windows. The windows were 400-ms blocks, with two separate blocks in the offer 1 window and two more blocks in the offer 2 window. The first blocks in a window were from 0 ms (offer on) to 400 ms after (offer off), and the second window was from 450 ms to 850 ms after the offer was presented. We refer to the latter window as the delay window. The estimated model had coefficients for subjective value, offer spatial position and their interaction, which can be used to estimate the nonlinear encoding terms. Model terms for offer side were effects coded [-1 1]; expected value was estimated with linear and b-spline terms (see Extended Data Fig. 2 for more details on the splines), with the value term being min–max normalized to standardize across units and monkeys. Therefore, the design matrix in total had five variables to estimate plus the intercept. The model formula for a neuron (*n*) firing rate (FR) was estimated using all trials:

$$\text{FR}(n, \text{trials}) = \beta_0 + \beta_1 \text{Value} + \beta_2 \text{Space} + \beta(\text{Interaction}) + N(0, \sigma^2)$$

Computing population encoding subspaces for space and value (Fig. 2). To measure separability between spatially distinct value subspaces, we used the computed regression coefficients for calculating the population encoding subspaces. Our goal was to compare the value subspaces for left and right offers both within offer epochs (for example, offer 1) and across offer epochs. Because the coefficients (β) were derived from a linear model, we combined the coefficients to create a value vector for each distinct side and epoch.

Computation of each subspace involves setting the levels of *X* in the βX of the regression equation, given a set of fit β s per neuron (see 'Single-neuron linear regression model' above). Essentially, this process gives the model predicted firing rate spanning from the lowest range normalized value offer that includes the intercept offset (0) to the highest (1). The process of creating the value subspace vector for the left side, for example, proceeded as follows: first, we must subtract two vectors, with the base vector centered on the different sides being compared (for example, left and right), as they themselves have different intercepts. Given the linear model form for the design matrix, and assuming left = 1 and right = -1, the intercept vector for value = 0 and left offer subspace, the vector is:

Value high vector : $[1]\beta_0 + [1]\beta_{\text{Value}} + [1]\beta_{\text{Space}} + [1]\beta_{\text{Interaction}}$

Value/Side/Intercept vector : $[1]\beta_0 + [0]\beta_{\text{Value}} + [1]\beta_{\text{Space}} + [0]\beta_{\text{Interaction}}$

In total, there were six distinct task-functional comparisons, including comparisons of (1) left and right within offers 1 and 2, (2) left and right between offers 1 and 2, and (3) between the same-side across offers (that is, left and left for offer 1 and 2, right and right for offer 1 and 2).

The testing of separability (or orthogonality) between subspaces requires establishing a null hypothesis of 1. This is because separable subspaces will be less correlated than a perfect correlation under the influence of noise. Specifically, because our main hypothesis was essentially a close to zero correlation between subspaces (that is, $|r| = 0$), we needed to estimate how a perfect correlation in the dataset, but confounded by noise, would be distributed ($|r| = 1$). Therefore, we consider a subspace as separable or (semi-)orthogonal if the correlation between neuron weights is outside the confidence bound of this hypothetical perfect correlation distribution.

We addressed this problem by applying an already established approach⁸¹. In brief, we fit the linear models described above on 1,000 bootstrap resampled sets of trials. For each set, we computed the subspace correlation between left-value and right-value representations. Then, to construct the noise ceiling estimate, we computed the subspace correlation between the left-value subspaces for the first and last 500 resamples (yielding 500 subspace correlation estimates) as well as the correlation between the right-value subspaces from the first and last 500 resamples (yielding another 500 estimates). We then compared the 1,000 subspace correlation estimates between the left-value and right-value subspaces with the 1,000 total estimates from the noise ceiling distribution using the bootstrap test. We also repeated this analysis pipeline on data where the actual value and side of each trial was shuffled relative to the neural activity. On each bootstrap iteration, we chose a different shuffle for a total of 1,000 shuffles. Then, we computed the subspace correlation as described above.

Finally, we repeated this same procedure for the b-spline value representation models, using the alignment index^{81,82} instead of the subspace correlation. These results are shown in Extended Data Fig. 2.

While we also performed a model comparison analysis, which gives insight into how many single neurons are best fit by both this model with a nonlinear interaction and a model without such an interaction, we compute all subspace correlations using this full nonlinear model. There are two reasons for this: First, the model with an interaction encapsulates the model without an interaction (that is, the interaction coefficient has zero strength) and our bootstrap procedure already provides a distribution of estimates of the value of the interaction coefficient. Along with our comparison to the estimate of the noise ceiling, this provides a reasonable estimate of the subspace correlation. Second, the fitting procedure used to estimate the ‘weight’ of each model type below does not have an established method for estimating the uncertainty of the weights. However, we still explore whether using the best fitting model would drastically change our results in Extended Data Fig. 3b. This analysis shows a qualitatively similar pattern of results, although there is a striking change to the noise floor for all regions.

Finally, in the following work, we obtain an even more stringent estimate of subspace correlation through our cross-validated distance estimation procedure (see ‘Linear–nonlinear data decomposition’). This qualitatively replicates our findings, although the difference between VS and the noise ceiling does not reach significance under this more stringent analysis.

Comparison between models with parallel and nonparallel subspaces (Fig. 2). To understand the single-neuron basis of our semi-orthogonal subspaces, we performed a model comparison analysis for three classes of models: First, a regression model that fit

only a noise term, and therefore explained the data only as random fluctuations. Second, models that had both side and value terms, but with no interaction between them. At the population level, this model would produce parallel subspaces. We fit these models with both linear and nonlinear representations of value (Extended Data Fig. 2). Third, models that have both side and value terms as well as an interaction term between them. If the interaction term is significant, then, at the population level, this model produces nonparallel subspaces.

For each model, we obtained a posterior predictive distribution, which is a distribution of samples from the model combined across all trials. We used this posterior predictive distribution to check the fit of our model, and we found that our models provide good fit to the value–response curves of the actual neurons (Fig. 2b and Extended Data Fig. 2B). Then, we performed a Bayesian model stacking analysis, which assigns a weight to each model⁸³. In the stacking analysis, the posterior predictive distributions of all the models are combined to produce a combined model (that is, a weighted sum of posterior predictive samples) that maximizes the leave-one-out cross-validation performance. We obtained similar results for more traditional measures of model goodness of fit, such as the widely applicable information criterion⁸⁴. However, we believe the stacking analysis provides useful intuition: It gives insight both into which model provides the best fit, as well as into which combination of models provides the best fit.

Single-neuron basis of subspaces: bimodality. To test whether the subspaces are primarily formed from gain modulating neurons, we note that the distribution of firing rate differences (across space) to value tuning should deviate from unimodal. We tested this by asking whether the distribution of differences in value tuning vectors for left and right offers was bimodal, using Hartigan’s dip test of bimodality. Specifically, we formed distributions of tuning differences using the left- and right-value subspace coefficients for each neuron computed above. The distributional responses were computed for each offer (1 and 2) and time window (offer on and delay) as a sensitivity index:

$$\text{Value – space FR sensitivity} = \frac{\text{FR}(n)_{\text{left}} - \mu(\text{FR}(n)_{\text{left}})}{\sigma(\text{FR}(n)_{\text{left}})} - \frac{\text{FR}(n)_{\text{right}} - \mu(\text{FR}(n)_{\text{right}})}{\sigma(\text{FR}(n)_{\text{right}})}$$

In the above, for example, $\text{FR}(n)_{\text{left}}$ is the value-left subspace vector coefficient for neuron n .

Single-neuron basis of subspaces: subspace contribution of gain modulation versus complex nonlinear. The bimodality test described above provides a global measure of whether the populations of neurons composing the subspaces are generally specialized to certain spatial locations. To further demarcate the contribution of gain versus complex nonlinear encodings to the subspaces, we used the property that gain modulated neurons will have the same preferred response to value, but with different amplitudes for different spatial positions, while complex nonlinear neurons will have a shifting preference to value. The main result we aimed for was determining the prevalence of nonlinear complex versus gain modulating neurons, and determining their percentage contribution of each type in forming the subspaces. To do this, we first determined which neurons in a region contributed to both left- and right-value subspaces for each epoch and time window. Subspace contribution was determined by finding each neuron’s percentage variance in each subspace:

$$\text{Subspace contribution (\%)} = \frac{(\text{FR}(n)_{\text{space}}^2)}{\sum (\text{FR}(n)_{\text{space}}^2)}$$

The above neuron participation ratio is inherently related to the dimensionality of that subspace; effectively, each neuron’s percentage

variance contributed can be viewed as an approximation to the eigenvalues of the subspace^{85,86} that retained the top 95% of neurons, which was an average of 48 neurons in each subspace, across all brain regions. We then compared whether each neuron's preferred value was the same in left and right subspaces, within each of the analysis windows used for the regressions. Specifically, we performed a bootstrap hypothesis test of mean differences. The mean firing rates for testing were resampled 1,000 times for each neuron, and the mean rate was computed over seven equally distributed levels of value, separately for left and right offers. The peak rate for each resample was collected in a vector and subsequently randomized 1,000 times (with resampling) to create a null distribution of mean differences in firing preference. The P value was computed by counting the number of times the randomly permuted mean difference was larger than the empirically estimated difference.

Binding by subspace orthogonalization

To motivate the problem we are studying here, we begin by considering a neural code that does not bind the distinct features of a single stimulus together. For example, a factorized representation of different stimulus features—even if the code for each individual feature is nonlinear—will fail to distinguish between different possible stimulus set bindings. In our setting, the neural population responds to both offer position and offer value. On a single trial the animal is shown a set of two stimuli, X , which each have a position and value. So, this set can be written as $X = \{[p_1, v_1], [p_2, v_2]\}$, where p_1 is the position of offer 1 and v_1 is its value. Then, if the average response of a neural population \bar{r} is given by a function $r(X) = \sum_x f(x)$, we can write

$$\bar{r} = f([p_1, v_1]) + f([p_2, v_2])$$

Now, if the function f can be factorized into a sum of functions, f_{pos} for offer position and f_{val} for offer value, then we can rewrite the response as

$$\bar{r} = f_{\text{val}}(v_1) + f_{\text{pos}}(p_1) + f_{\text{val}}(v_2) + f_{\text{pos}}(p_2)$$

This response contains all the original information about the features of the two offers—but it does not preserve their binding. In particular, while a maximum likelihood decoder would have a maximum at the correct stimulus set X , it would also have an equivalent maximum at the chimeric stimulus set $S = \{[p_1, v_2], [p_2, v_1]\}$. This is because the average population response to the two stimulus sets is exactly the same,

$$\begin{aligned} \Delta &= r(X) - r(S) \\ &= f_{\text{val}}(v_1) + f_{\text{pos}}(p_1) + f_{\text{val}}(v_2) + f_{\text{pos}}(p_2) \\ &\quad - f_{\text{val}}(v_2) - f_{\text{pos}}(p_1) - f_{\text{val}}(v_1) - f_{\text{pos}}(p_2) \\ &= 0 \end{aligned}$$

Thus, in the case of a high-value offer on the left and a low-value offer on the right, an alternative and equally likely interpretation of the neural representation would be that there was a low-value offer on the left and a high-value offer on the right, which would lead to a suboptimal choice by the animal. Even further, for a linear representation of value and position (that is, f_{val} only multiplies v by a scalar), there would be many more equally likely chimeric stimulus sets, including a set with only the single stimulus $[p_1 + p_2, v_1 + v_2]$ and any set of stimuli with values and positions that have the same sum as the presented stimulus set. These additional ambiguities can be resolved through nonlinear (but still separate) representations of position and value as well as through the inclusion of side information about the number of stimuli in the set in the decoder. Such side information could be provided by brain regions shown to represent the number of stimuli in a scene, such as lateral prefrontal cortex and posterior parietal cortex⁸⁷. Here, we focus on binding errors and decoding with side information about the number of stimuli.

In practice, humans and other animals may make these kinds of binding errors, but they are thought to be infrequent⁵³. So, in many cases, this ambiguity must be successfully resolved. To understand how this happens, we consider the case when f cannot be fully factorized and instead can be written as $f([p, v]) = f_{\text{pos}}(p) + f_{\text{val}}(v) + f_{\text{pos-val}}([p, v])$, where $f_{\text{pos-val}}$ is a non-factorizable function of both side and value. We refer to this term as the conjunctive part of the response. Then, the difference in response for the correct and chimeric stimulus sets Δ from before can be written as,

$$\begin{aligned} \Delta &= r(X) - r(S) \\ &= f_{\text{pos-val}}([v_1, p_1]) + f_{\text{pos-val}}([v_2, p_2]) - f_{\text{pos-val}}([v_2, p_1]) - f_{\text{pos-val}}([v_1, p_2]) \end{aligned}$$

Thus, so long as the squared sum of Δ across the neural population is larger than zero—and, in a noisy system, larger than the noise—then a decoder will be able to resolve the ambiguity between the correct and chimeric stimulus set and avoid misbinding errors. We derive an expression for this misbinding rate in a simplified case below. One of our key theoretical results is that the inclusion of this conjunctive part of the representation can simultaneously resolve the coding ambiguity without destroying the benefits of a factorized representation.

At the level of single neurons, the conjunctive part of the representation $f_{\text{pos-val}}$ manifests as neurons with nonlinear mixed selectivity for offer value and position. At the population level, the conjunctive part of the representation manifests as orthogonalization of the subspaces encoding the value of left and right offers.

Linear–nonlinear code framework (Figs. 3 and 4)

To apply our mathematical framework to the experimental data, we considered a discretized offer value along with offer position—value is discretized as for our decoding analysis (see below). In these data, we found that decoding the binarized value yielded much better performance than decoding the continuous value with various methods. The analytic approach we take below is also simplified for discrete value. However, the subspace binding hypothesis does not depend on this discreteness, as discussed above. Once we discretize value, we have $K = 2$ latent variables that each take on $n = 2$ different values. However, the theory we develop applies to any choice of K and n , and so can be applied well beyond this experimental setting.

We model the neural responses of N neurons as,

$$\mathbf{r}(x) = L\mathbf{x}_z + Mf_N(x) + \mathbf{e}$$

where $\mathbf{r}(x)$ is an $N \times 1$ vector of the z-scored activity of N neurons, L is an $N \times K$ linear transformation of the z-scored stimulus vector \mathbf{x}_z and has columns $L = [d_{LV}L_1 \ d_{LA}L_2]$, M is an $N \times n^K$ linear transformation of $f_N(x)$, which is a nonlinear transformation of the stimulus vector \mathbf{x} (defined in detail below). Finally, $\mathbf{e} \sim \mathcal{N}(0, \sigma^2)$. We z-scored the stimulus vector \mathbf{x} , so that the linear distance in the representation produced after the linear transform is controlled fully by the linear transform L , and does not depend on the specific choice of coding for \mathbf{x} .

In the experiment, there were two features that each took on two values. So, \mathbf{x} is a vector with two elements that are each either 1 or 2. The first element corresponds to low value (1) or high value (2); the second element corresponds to the two values of offer position. The possible \mathbf{x} are:

$$\begin{aligned} \mathbf{x}_{11} &= [11]^T \\ \mathbf{x}_{12} &= [12]^T \\ \mathbf{x}_{21} &= [21]^T \\ \mathbf{x}_{22} &= [22]^T \end{aligned}$$

The nonlinear function we consider is the conjunctive identity function used in refs. 20,24, where

$$f_N(x)_{ij} = [x_i = i][x_j = j]$$

and

$$f_N(x) = [f_N(x)_{11} f_N(x)_{12} f_N(x)_{21} f_N(x)_{22}]^T$$

For the linear transform L , which has $K = 2$ columns, the length of the first column is d_{LV} , which will be the average distance between two stimulus representations that differ only in their value (where $M = 0$). We will refer to the length of the second column as d_{LA} , the average distance between two stimuli that differ only in offer position (also where $M = 0$). All the columns of M will have the same length m , which will mean that the distance between the nonlinear components of two stimulus representations will be $d_N = \sqrt{2}m$. We assume that the length of the nonlinear perturbation for each stimulus is the same; while this is unlikely to be precisely true, it simplifies our analysis and still gives good results when comparing to the experimental data. We also assume that the columns of both L and M are orthogonal, both within each matrix and between the two matrices (although the analytic results are similar without the between matrix orthogonality). For large N , this will tend to be true for random vectors.

Finally, a code in this framework is described by a stimulus set (K and n) and four parameters: d_{LV} (representing offer value), d_{LA} (representing offer position), d_N (representing the nonlinear part of the code) and σ (representing the standard deviation of the noise). In our analytic theory, we show that the binding error rate depends on d_N and σ . That is, low binding error rates can be achieved so long as d_N is sufficiently large, regardless of the strength of the linear part of the code (d_{LV} or d_{LA}). We also show that the generalization error rate depends only on d_{LV} , d_N and σ – that is, it does not depend on d_{LA} . Thus, we will focus on estimating d_{LV} , d_N and σ from the experimental data.

Linear–nonlinear code with constrained power. To compare different coding strategies with each other, we will fix the total amount of spiking ‘power’ available to the code, while varying the trade-off between the allocation of the power to the linear or nonlinear part of the code. We also use this constrained power to compare the representational geometry observed in different brain regions (Fig. 4). This spiking power is the summed variance across all units in the code, which is often used as a proxy for the metabolic energy consumption of the code²⁴. We keep the total power P constant and vary the linear power P_L and nonlinear power P_{NL} . In particular, $P = P_L + P_{NL}$. To understand how changes to the relative linear and nonlinear power of the code shape the representational geometry, we begin by deriving how linear and nonlinear power translate to the linear and nonlinear distances in representational space introduced above.

Linear distance. We derive the distance between adjacent stimuli in the linear code for a particular number of features K and number of values n that each feature takes on as well as the linear power of the code (P_L),

$$d_L = \sqrt{\frac{12P_L}{K(n^2 - 1)}}$$

We approach this by computing the variance (that is, the linear power P_L) of a uniformly sampled K -dimensional lattice with n points spaced at distance d_L along each dimension. Then, we invert the expression for the variance to find an expression for the distance between the points. First, we write the variance P_L as

$$\begin{aligned} n^K P_L &= \sum_{i=0}^{n-1} \left[\left(i - \frac{n-1}{2} \right)^2 d_L^2 + \sum_{j=0}^{n-1} \left[\left(j - \frac{n-1}{2} \right)^2 d_L^2 + \dots \right] \right] \\ &= \sum_i^{n-1} \sum_j^{n-1} \dots \sum_k^{n-1} \left(i - \frac{n-1}{2} \right)^2 d_L^2 + \left(j - \frac{n-1}{2} \right)^2 d_L^2 + \dots + \left(k - \frac{n-1}{2} \right)^2 d_L^2 \\ &= K n^{K-1} \sum_i^{n-1} \left(i - \frac{n-1}{2} \right)^2 d_L^2 \\ &= K n^{K-1} d_L^2 \sum_i^{n-1} i^2 - (n-1) \sum_i^{n-1} i + n \frac{n-1}{2} \end{aligned}$$

and we can rewrite this with known expressions for the sum of integers and sum of squared integers up to a particular value,

$$\begin{aligned} n^K P_L &= K n^{K-1} d_L^2 \left[\frac{(n-1)n(2n-1)}{6} - \frac{n(n-1)^2}{2} + \frac{n(n-1)^2}{4} \right] \\ &= K n^K d_L^2 \left[\frac{(n-1)(2n-1)}{6} - \frac{(n-1)^2}{4} \right] \\ &= K n^K d_L^2 \left[\frac{2n^2 - 3n + 1}{6} - \frac{n^2 - 2n + 1}{4} \right] \\ &= K n^K d_L^2 \left[\frac{4n^2 - 6n + 2}{12} - \frac{3n^2 - 6n + 3}{12} \right] \\ n^K P_L &= K n^K d_L^2 \frac{n^2 - 1}{12} \\ P_L &= K d_L^2 \frac{n^2 - 1}{12} \end{aligned}$$

Now, we rewrite in terms of d_L ,

$$d_L = \sqrt{\frac{12P_L}{K(n^2 - 1)}}$$

which is the expression given above.

Following from the lattice structure, stimuli at a diagonal point on the lattice have distance $\sqrt{2}d_L$.

Linear neighbors derivation. Second, we find the average number of neighbors that a particular stimulus has at both this nearest distance N_{LA} and the nearest diagonal distance N_{LD} . In the experiments here, where there are $K = 2$ features that each take on $n = 2$ values, the number of nearest neighbors in the linear code is two for every point, and the number of neighbors at the nearest diagonal distance is one. This encompasses all the stimuli. However, in general, this is a counting problem. We observe that, in the lattice, there are two edge values for each feature and $n - 2$ non-edge values. Thus,

$$N_{LA} = \frac{1}{n^K} \sum_{c=0}^K (2K - c) \binom{K}{c} (n - 2)^{K-c} 2^c$$

and

$$N_{LD} = \frac{1}{n^K} \sum_{c=0}^K \left(4 \binom{K-c}{2} + 2(K-c)c + \binom{C}{2} \right) \binom{K}{c} (n-2)^{K-c} 2^c$$

Nonlinear distance. The nonlinear distance is

$$d_N = \sqrt{2P_N}$$

and has been treated in more detail previously²⁴.

Nonlinear neighbors derivation. Because each nonlinear representation is along a vector that is orthogonal to all other nonlinear

representations, from a particular stimulus all other representations are at minimum distance. So,

$$N_{NL} = n^K - 1$$

Total code distance. The total code distance for orthogonal linear and nonlinear code parts is

$$d_C = \sqrt{d_L^2 + d_{NL}^2}$$

If the linear and nonlinear code parts are not constrained to be orthogonal, then the total code distance is a random variable with the following form,

$$d_C = \sqrt{d_L^2 + d_{NL}^2 + 2d_{NL}d_L\eta}$$

where $\eta \sim \mathcal{N}(0, 1/M)$ due to the fact that the dot product of two unit vectors is normally distributed with variance inverse to their length (that is, the distribution of η). The distance is similarly defined for the next-nearest stimuli,

$$d_{C+1} = \sqrt{2d_L^2 + d_{NL}^2 + \sqrt{8}d_{NL}d_L\eta}$$

Total code neighbors. To combine the code neighbors, it is enough to simply take the minimum between the linear and nonlinear parts (assuming that the linear and nonlinear parts are both nonzero), which will always be equal to N_{LA} (or N_{LD} for next-nearest). So,

$$N_C = N_{LA}$$

$$N_{C+1} = N_{LD}$$

Relating linear and nonlinear distance to subspace correlation. The linear–nonlinear code framework developed here provides a different way to define the subspace correlation measured used in the rest of the paper. We compute subspace correlation directly from the linear (d_{LV}) and nonlinear distances (d_N). In particular, we can take the cosine similarity between \mathbf{v}_l and \mathbf{v}_r , as defined in the main text,

$$\rho = \frac{\mathbf{v}_l \cdot \mathbf{v}_r}{|\mathbf{v}_l|_2 |\mathbf{v}_r|_2}$$

$$= \frac{(L_1 + M_1 - M_2) \cdot (L_1 + M_3 - M_4)}{d_{LV}^2 + d_N^2}$$

$$= \frac{d_{LV}^2}{d_{LV}^2 + d_N^2}$$

where ρ is the subspace correlation (see Fig. 3a for a schematic).

The overall error rate of a linear–nonlinear code. Using the expressions developed in the previous sections, we can write an approximation for the likelihood that a linear–nonlinear code makes an error, for a given stimulus set, defined by K and n , and given linear and nonlinear powers P_L and P_{NL} .

Here, we define an error as the most likely stimulus (set) under a maximum likelihood decoder \hat{x} not being the original stimulus (set) x . We take the a nearest-neighbor union bound approach^{24,88,89} and develop the following expression for the error rate,

$$P(\text{error}) \approx N_C Q\left(-\frac{d_C}{2\sigma}\right) + N_{C+1} Q\left(-\frac{d_{C+1}}{2\sigma}\right)$$

Where $Q(\cdot)$ is the standard Gaussian cumulative distribution function. Thus, we can see that the error rate depends most strongly on the distances. From our distance definitions, we know that, if $K > 2$ or $n > 2$, increasing nonlinear power is the most efficient way to increase

distance. As a consequence, to drive the error rate down, it is best to put all code power toward the nonlinear part.

For multiple stimuli, the code error rate can be written as,

$$P(\text{error}) \approx SN_C Q\left(-\frac{d_C}{2\sigma}\right) + SN_{C+1} Q\left(-\frac{d_{C+1}}{2\sigma}\right) + P(\text{binding error})$$

where S is the number of stimuli. We will develop an approximation for $P(\text{binding error})$ next.

Derivation of the binding error rate. We begin by considering a purely linear code for multiple stimuli X , where

$$r_L(X) = \sum_{\mathbf{x} \in X} L\mathbf{x}_z + \epsilon$$

where \mathbf{x}_z is the z-scored features of \mathbf{x} and L is an $N \times K$ linear transform—as above. With a purely linear code and stimuli that are defined by two features $K = 2$ that each take on two values $n = 2$, there are two stimulus pairs that give rise to exactly the same average response. Those pairs are

$$X = \{\mathbf{x}_{11}, \mathbf{x}_{22}\}$$

$$S = \{\mathbf{x}_{12}, \mathbf{x}_{21}\}$$

which is to say,

$$\bar{r}_L(X) = \bar{r}_L(S)$$

Due to this property, even a decoder that is optimal for our current setting (the maximum likelihood decoder), will not be able to discriminate between these two options. If X is presented, then we refer to trials in which S is decoded from the activity by a maximum likelihood decoder as a misbinding error. For this purely linear code, roughly half the time X is presented, S will be decoded. How can we modify the code to make fewer misbinding errors? Here, we show that reintroducing the nonlinear part of the code can effectively drive down the probability of binding errors, precisely by increasing the distance between the representations of these two stimulus sets.

Now, we include the nonlinear term in the code, as discussed above. So, we want to show how increasing the nonlinear distance d_N decreases the probability of misbinding errors. First, we derive the distance between the average representation of X and S in the full code (that is, with nonzero nonlinear distance), where,

$$r(X) = \sum_{\mathbf{x} \in X} L\mathbf{x}_z + Mf_N(\mathbf{x}) + \epsilon$$

So, for X and S defined above,

$$d_S = |\bar{r}(X) - \bar{r}(S)|_2$$

$$= \left| \sum_{\mathbf{x} \in X} L\mathbf{x}_z + Mf_N(\mathbf{x}) - \sum_{\mathbf{s} \in S} L\mathbf{s}_z - Mf_N(\mathbf{s}) \right|_2$$

$$= \left| \sum_{\mathbf{x} \in X} Mf_N(\mathbf{x}) - \sum_{\mathbf{s} \in S} Mf_N(\mathbf{s}) \right|_2$$

$$= |M_1 + M_2 - M_3 - M_4|_2$$

$$= \sqrt{2}d_N$$

We notice that this does not depend on the particular stimulus sets X and S anymore. Indeed, due to our assumption that the nonlinear distances are constant across stimuli, this is the distance between any two sets of stimuli that are linearly confusable. As anticipated by the above example, the distance d_S does not depend on the linear distance d_L , only on the nonlinear distance d_N . This is because the linear part of the code is ambiguous with respect to binding (as explained in

‘Binding by subspace orthogonalization’ above), so making it larger will not help resolve confusion about which set of stimuli was seen.

Now, knowing the distance between the correct stimulus set X and a misbound stimulus set S , d_S , we can use the following expression for the rate of binding errors via a union bound approximation where S is a stimulus set that is linearly confusable with X , C_X is the set of all stimulus sets that are linearly confusable with X , and \hat{X} is the stimulus set inferred by a maximum likelihood decoder:

$$\begin{aligned} P(\hat{X} \in C_X | r(X)) &= P\left(\bigcup_{S \in C_X} \hat{X} = S | r(X)\right) \\ &\leq \sum_{S \in C_X} P(\hat{X} = S | r(X)) \\ &= \sum_{S \in C_X} P(d(\bar{r}(S), r(X)) < d(\bar{r}(X), r(X))) \\ &\approx \sum_{S \in C_X} Q\left(-\frac{d_S}{2\sigma}\right) \end{aligned}$$

where $d(\cdot, \cdot)$ takes the Euclidean distance between its two arguments and $Q(\cdot)$ is the standard Gaussian cumulative distribution function. Then, we average over X ,

$$P(\text{binding error}) \approx N_S Q\left(-\frac{d_S}{2\sigma}\right)$$

where N_S is the average size of C_X across all X . In our case, for two stimuli with two features that each take on two values, $N_S = \frac{1}{4}$. In general, we can write the number of chimeric stimulus sets as

$$N_S = \frac{1}{2} \binom{S}{2} \sum_{i=0}^{K-1} \binom{K}{i} \left(\frac{1}{n}\right)^i \left(1 - \frac{1}{n}\right)^{K-i} \sum_{j=1}^{K-i} \binom{K-i}{j}$$

Derivation of the generalization error rate. Next, we want to develop a prediction for the generalization error rate that depends on our four parameters d_{LV} , d_{Ne} and σ as well as on the s.e.m. ϵ . In particular, we can compare the generalization error rate predicted by our theory and developed here from the generalization error rate of decoders trained on the neural data (discussed more below). This provides a crucial validation test for our theory, and the close correspondence observed between the theory and the data indicate that our formalization captures the relevant aspects of the geometry of the neural representations.

Here, we will develop the approximation assuming that the linear and nonlinear parts of the code are orthogonal to each other. However, the results are similar if the linear and nonlinear parts are randomly chosen with respect to each other.

As before, we have four stimuli of interest x_{ij} for $i, j \in \{1, 2\}$. We can write the representation corresponding to each of them in terms of the linear and nonlinear code components where M_i denotes the columns of M and L_i denotes the columns of L , such that,

$$\begin{aligned} r(\mathbf{x}_{11}) &= 0 \\ r(\mathbf{x}_{21}) &= d_{LV}L_1 + d_{Ne}M_{12} \\ r(\mathbf{x}_{12}) &= d_{LA}L_2 + d_{Ne}M_{13} \\ r(\mathbf{x}_{22}) &= d_{LV}L_1 + d_{LA}L_2 + d_{Ne}M_{14} \end{aligned}$$

where d_{LA} is the linear distance associated with the variable that is being generalized across (that is, offer side or time), d_{Ne} is the combined distance from both the nonlinear part of the code and the s.e.m. $d_{Ne} = \sqrt{d_N^2 + \epsilon^2}$, and we have defined

$$M_{ij} = \frac{M_i - M_j}{\sqrt{2}}$$

for convenience. The standard error distance ϵ appears here because some of the perturbations to the underlying linear structure are reliable and form the nonlinear distance, while others are unreliable and emerge due to the noisy estimation of each centroid. The generalization performance of the classifier is reduced by both—while, for instance, the traditional cross-validated performance of the classifier would be reduced only by the s.e.m.

In this approach, we take the perspective of a linear prototype decoder (similar to the one analyzed in ref. 45). During training, this decoder sees only two distinct stimuli (high-value and low-value offers presented at a single position) and learns to distinguish them from each other. In particular, it learns a prototype representation for each stimulus (the average across many noisy presentations) and then decides the category of a new stimulus by evaluating which of the two prototypes the new stimulus is closer to. In practice, this is done by separating the two classes with a linear hyperplane that is defined by the vector that connects the two class prototypes. Since the prototype decoder evaluates new stimuli based only on this vector, we can then apply this learned vector to novel stimulus classes and examine the statistics of these novel representations in the single dimension that is relevant to decoding. We show that the generalization performance predicted by our analysis of the relatively simple prototype decoder matches the performance of a more sophisticated support vector machine when applied to real data.

In our framework, we can write the vector that defines the decoding hyperplane given two training stimuli \mathbf{x}_{11} and \mathbf{x}_{21} ,

$$\begin{aligned} \mathbf{v}_D &= \frac{1}{c} (r(\mathbf{x}_{21}) - r(\mathbf{x}_{11})) \\ &= \frac{1}{c} (d_{LV}L_1 + d_{Ne}M_{12}) \end{aligned}$$

where c normalizes \mathbf{v}_D to be a unit vector and has the form

$$c = \sqrt{d_{LV}^2 + d_{Ne}^2}$$

Because we assume that the noise magnitude is the same for each stimulus, the decoding boundary is $\frac{c}{2}$. To decode an unseen data point y , we would evaluate,

$$o = \text{sgn}\left[\mathbf{v}_D \cdot y - \frac{c}{2}\right]$$

where ‘sgn’ takes the sign of its argument. If $o = -1$, then we would classify y as, for instance, low value; if $o = 1$, then we would classify y as high value.

So, to evaluate the generalization performance of this decoder on held-out stimulus conditions, we can apply this same logic to the left-out stimuli \mathbf{x}_{12} and \mathbf{x}_{22} . In particular, we want to project the representations of these held-out stimuli onto the decoder vector \mathbf{v}_D derived above and then compare the position of the representations along that vector to the decoding threshold $\frac{c}{2}$. So, for \mathbf{x}_{12} ,

$$\begin{aligned} d_{12} &= \mathbf{v}_D \cdot r(\mathbf{x}_{12}) - \frac{c}{2} \\ &= \frac{1}{c} (d_{LV}L_1 + d_{Ne}M_{12}) \times (d_{LA}L_2 + d_{Ne}M_{13}) - \frac{c}{2} \\ &= \frac{1}{c^2} d_{Ne}^2 - \frac{c}{2} \\ &= \frac{1}{c^2} d_{Ne}^2 - \frac{c^2}{c^2} \\ &= \frac{d_{Ne}^2}{c^2} - \frac{d_{LV}^2 + d_{Ne}^2}{c^2} \\ &= -\frac{\frac{1}{2}d_{LV}^2}{\sqrt{d_{LV}^2 + d_{Ne}^2}} \end{aligned}$$

and, for x_{22} ,

$$\begin{aligned} d_{22} &= \mathbf{v}_D \cdot r(\mathbf{x}_{22}) - \frac{c}{2} \\ &= \frac{1}{c} (d_{LV}L_1 + d_{Ne}M_{12}) \times (d_{LA}L_2 + d_{LV}L_1 + d_{Ne}M_{14}) - \frac{c}{2} \\ &= \frac{1}{c} \left(d_{LV}^2 + \frac{1}{2}d_{Ne}^2 \right) - \frac{c}{2} \\ &= \frac{\frac{1}{2}d_{LV}^2}{\sqrt{d_{LV}^2 + d_{Ne}^2}} \end{aligned}$$

Now, using these two distances and the noise magnitude σ , we can predict how well a decoder trained to discriminate \mathbf{x}_{11} from \mathbf{x}_{21} will generalize to discriminate \mathbf{x}_{12} from \mathbf{x}_{22} . This is the cross-category generalization performance from ref. 27 and that is discussed in the main text. In particular, the

$$\begin{aligned} P(\text{CCGP error}) &\approx \frac{1}{2}Q\left(\frac{d_{12}}{\sigma}\right) + \frac{1}{2}Q\left(-\frac{d_{22}}{\sigma}\right) \\ &= Q\left(-\frac{1}{\sigma} \frac{\frac{1}{2}d_{LV}^2}{\sqrt{d_{LV}^2 + d_{Ne}^2}}\right) \end{aligned}$$

where Q is the cumulative distribution function of a standard normal distribution, as before.

Decoding analyses (Figs. 4 and 5)

Binarizing value. We discretize value into high and low by splitting it according to the subjective value transformation computed for each session and excluding the middle 15th percentile.

Preprocessing the neural data. Before decoding, we preprocessed the data by z-scoring and then applying a principal component analysis that retains enough dimensions to capture 99% of the variance. Both the z-score and principal component analysis transforms are fit on the training set only.

Across-time decoding. Across-time decoding analyses are done on data split into 500-ms moving averages with a step of 20 ms between bins. Decoding models are fit independently for each timestep.

Time-pooled decoding. Pooled decoding analyses are done on data from three (Fig. 4) or four (Figs. 5 and 6) nonoverlapping 300-ms bins that begin 100 ms after offer onset. The activity from each neuron in each time bin are treated as separate features and all used for decoding in a single model. All of the bins from a single trial are in either the training or the testing set; they are not split across both.

For decoding, we consider value presented in two distinct conditions. We begin by constructing pseudopopulations for high value and low value in each condition (for example, high value and low value on the left and high value and low value on the right). The pseudopopulation consists of all neurons from a particular brain region with at least 160 trials for each of the four conditions for the first set of decoding and distance estimation analyses (Fig. 4; for example, splitting into left and right presentations with high or low value, combining across offer 1 and offer 2) and at least 40 trials for the second set of distance estimation analyses (Fig. 6; for example, splitting into all eight distinct experimental conditions). These criteria were chosen to maximize the number of both trials and neurons that could be included in the pseudopopulation analyses (see Extended Data Fig. 10 for the dependence of included neurons on required trials). Our results are stable for other reasonable choices of this cutoff, but become substantially noisier as fewer trials are required or neurons are included.

Decoding. Then, we trained a support vector machine decoder with a linear kernel to discriminate high value from low value in one condition

(for example, decoding value from only presentation on the left) and tested that decoder both on held-out trials from that condition (10% of trials are held out) and on all trials from the other condition (for example, high value and low value on the right). The performance of the decoder on the held-out trials from the training condition is the standard decoding performance, and the performance of the decoder on the trials from the second condition is the cross-condition generalization performance.

All decoding analyses are implemented in scikit-learn⁹⁰.

In the main text, we compared the generalization performance of these support vector machine decoders to the generalization performance that is predicted by our analysis of the linear–nonlinear code model, given above. We believe that the close correspondence observed between the empirical and predicted generalization performance indicates that our linear–nonlinear code formalization captures relevant aspects of the neural representation geometry.

Estimating distances (Figs. 4 and 6)

To estimate both the linear and nonlinear distances used in our theory (Fig. 4) as well as the distance between conditions in the full representation space (Fig. 6), we use a cross-validated distance measure, often referred to as the crossnobis distance (although we did not incorporate an estimate of the noise structure into our use of the measure, so it is not a version of the Mahalanobis distance as the name crossnobis suggests), to estimate the unbiased Euclidean distance between every pair of stimulus conditions⁹¹. We used the routines provided by the Python RSA toolbox⁹². This yields a representational dissimilarity matrix⁹³. These matrices are symmetric and have as many rows and columns as there are distance experimental conditions. In the analyses in Fig. 4, there are four conditions. In the analysis in Fig. 6, there are eight. Each entry in the matrix gives the estimated distance between the conditions corresponding to the entry's row and column. We computed these matrices separately for each region, using data that are organized and preprocessed in the same way as for the decoding analyses described above.

Linear–nonlinear data decomposition (Fig. 4)

We want to estimate the three parameters of our code model, d_{LV} (the linear code distance for value), d_N (the nonlinear code distance) and σ (the noise standard deviation), from the data.

In particular, we can frame this as a constrained least-squares optimization problem. We want to find the solution to

$$\mathbf{d}_{\text{empirical}} = A\mathbf{d}_{\text{est}}$$

where $\mathbf{d}_{\text{empirical}}$ is a flattened version of the crossnobis distance matrix estimated above (now a 6×1 vector), A is a 6×3 design matrix where each row gives the integer multiples of the linear and nonlinear distances between the two corresponding conditions (for instance, conditions that have the same value, but different positions would have a row vector [0 1 2] in the matrix, while conditions with different values and different positions would have [1 1 2]), and $\mathbf{d}_{\text{est}} = [d_{LV}^2, d_{LA}^2, d_N^2]^T$. Then, we find the least-squares solution to the above equation with \mathbf{d}_{est} constrained to be nonnegative. See Extended Data Fig. 4 for this procedure applied to synthetic data, with a similar number of units and noise level as in our recordings—it accurately recovers both the linear and nonlinear distances.

Next, we estimate the magnitude of the noise in the neural representations. Here, we choose to estimate it along a single dimension that is particularly relevant for our analyses. For our generalization analysis, \mathbf{x}_{11} and \mathbf{x}_{21} are defined as the training set, while x_{12} and x_{22} are defined as the testing set. That is, the decoder will be trained to discriminate between $r(\mathbf{x}_{11})$ and $r(\mathbf{x}_{21})$ and then tested on its ability to discriminate between $r(\mathbf{x}_{12})$ and $r(\mathbf{x}_{22})$. Due to this, noise specifically along this learned decoding dimension, \mathbf{v}_1 , is most relevant

to generalization performance, and we will estimate specifically the magnitude of this noise from our data. In particular,

$$\sigma^2 = \mathbb{E}_{x_{ij}} \left[\frac{v_i}{|v_i|_2} \cdot (r(\mathbf{x}_{ij}) - \hat{r}(\mathbf{x}_{ij})) \right]^2$$

where the expectation is taken across both all trials from a particular condition and all conditions defined by i and j .

Finally, we also compute the s.e.m. of each of our estimates, collapsed into a standard error distance ϵ , which affects generalization performance as described above.

Normalizing representational power. In Fig. 4f, we normalized the representational power across regions to make a direct comparison between them. To do this, we take the estimated linear and nonlinear distances from the above procedure. Then, we use them to compute the representational power of the region, which is the sum of the powers in the linear and nonlinear parts of the code ('Linear–nonlinear code with constrained power'). Then, we normalize this power to a fixed value across all the regions (in Fig. 4f, the fixed power is equal to 10), before propagating this change backward to the estimated distances—which amounts to multiplying the linear and nonlinear distances by a region-dependent factor. As shown in Fig. 4f, this puts all the regions on the same spectrum, where if two regions have the same balance of linear and nonlinear distance (that is, subspace correlation), then they will also have the same predicted binding and generalization error rates. This was not true before normalization, because two regions could have the same balance but different total powers (Fig. 4f, inset).

In Fig. 6, we performed a similar normalization technique. However, because we did not estimate linear and nonlinear distance separately in that case, we normalized the representational power directly by embedding the distance matrix in Euclidean space using multi-dimensional scaling. Then, we normalized the power (that is, the L2 norm) of this representation, and propagated the change back into the distance matrix. This preserves the geometry of the representation while allowing us to independently manipulate the representational scale. The methods used in Figs. 4 and 6 are equivalent.

Further, we note that these normalizations of distance cannot change whether or not a particular distance is significantly different from zero—because the scaling is computed once and then applied uniformly to all resamples. Thus, it is useful for comparison purposes, but does not inflate our confidence or reduce our uncertainty about the underlying distances.

Statistics and reproducibility

No statistical method was used to predetermine sample size, but we adhered to customs in previous publications^{37,38,42}. Data were excluded from specific analyses as described in other sections of the methods; no data was excluded otherwise. The experiments were fully within subjects, and trials were presented in a random order. The investigators were not blinded to conditions during either the experiments or the analysis.

Reporting summary

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

Data availability

The raw data analyzed in this work are fully available on Figshare via <https://doi.org/10.6084/m9.figshare.26065600> (ref. 94).

Code availability

The code underlying this work relies on the Python scientific computing environment, including: python (3.8.2), numpy (1.23.5), scipy (1.10.1), sklearn (1.3.0), rsatoolbox and matplotlib (3.7.2). The custom

code written to generate the figures and analyze the data is available at https://github.com/wj2/subspace_binding. The version of the code used to generate the figures is available on Zenodo via <https://doi.org/10.5281/zenodo.12194146> (ref. 95).

References

- Paxinos, G., Petrides, M. & Evrard, H. C. *The Rhesus Monkey Brain in Stereotaxic Coordinates* (Academic Press, 2009).
- Blanchard, T. C., Hayden, B. Y. & Bromberg-Martin, E. S. Orbitofrontal cortex uses distinct codes for different choice attributes in decisions motivated by curiosity. *Neuron* **85**, 602–614 (2015).
- Heilbronner, S. R. & Hayden, B. Y. Dorsal anterior cingulate cortex: a bottom-up view. *Annu. Rev. Neurosci.* **39**, 149–170 (2016).
- Wang, M. Z., Hayden, B. Y. & Heilbronner, S. R. A structural and functional subdivision in central orbitofrontal cortex. *Nat. Commun.* **13**, 3623 (2022).
- Azab, H. & Hayden, B. Y. Correlates of decisional dynamics in the dorsal anterior cingulate cortex. *PLoS Biol.* **15**, e2003091 (2017).
- Sleezer, B. J., Castagno, M. D. & Hayden, B. Y. Rule encoding in orbitofrontal cortex and striatum guides selection. *J. Neurosci.* **36**, 11223–11237 (2016).
- Blanchard, T. C. & Hayden, B. Y. Neurons in dorsal anterior cingulate cortex signal postdecisional variables in a foraging task. *J. Neurosci.* **34**, 646–655 (2014).
- Blanchard, T. C., Wolfe, L. S., Vlaev, I., Winston, J. S. & Hayden, B. Y. Biases in preferences for sequences of outcomes in monkeys. *Cognition* **130**, 289–299 (2014).
- Heilbronner, S. R. & Hayden, B. Y. The description–experience gap in risky choice in nonhuman primates. *Psychon. Bull. Rev.* **23**, 593–600 (2016).
- Ebitz, R. B., Sleezer, B. J., Jedema, H. P., Bradberry, C. W. & Hayden, B. Y. Tonic exploration governs both flexibility and lapses. *PLoS Comput. Biol.* **15**, e1007475 (2019).
- Ferro, D., Cash-Padgett, T., Wang, M. Z., Hayden, B. & Moreno-Bote, R. Gaze-centered gating and re-activation of value encoding in orbitofrontal cortex. *Nat. Commun.* <https://doi.org/10.1101/2023.04.20.537677> (2023).
- Daunizeau, J., Adam, V. & Rigoux, L. VBA: a probabilistic treatment of nonlinear models for neurobiological and behavioural data. *PLoS Comput. Biol.* **10**, e1003441 (2014).
- Kimmel, D. L., Elsayed, G. F., Cunningham, J. P. & Newsome, W. T. Value and choice as separable and stable representations in orbitofrontal cortex. *Nat. Commun.* **11**, 3466 (2020).
- Elsayed, G. F., Lara, A. H., Kaufman, M. T., Churchland, M. M. & Cunningham, J. P. Reorganization between preparatory and movement population responses in motor cortex. *Nat. Commun.* **7**, 13239 (2016).
- Yao, Y., Vehtari, A., Simpson, D. & Gelman, A. Using stacking to average bayesian predictive distributions (with discussion). *Bayesian Anal.* **13**, 917–1007 (2018).
- Watanabe, S. A widely applicable Bayesian information criterion. *J. Mach. Learn. Res.* **14**, 867–897 (2013).
- Gao, P. et al. A theory of multineuronal dimensionality, dynamics and measurement. Preprint at *BioRxiv* <https://doi.org/10.1101/214262> (2017).
- Xie, Y. et al. Geometry of sequence working memory in macaque prefrontal cortex. *Science* **375**, 632–639 (2022).
- Nieder, A. The neuronal code for number. *Nat. Rev. Neurosci.* **17**, 366–382 (2016).
- Kim, J. H. J., Fiete, I. & Schwab, D. J. Superlinear precision and memory in simple population codes. Preprint at <https://arxiv.org/abs/2008.00629> (2020).
- Johnston, W. J. & Freedman, D. J. Redundant representations are required to disambiguate simultaneously presented complex stimuli. *PLoS Comput. Biol.* **19**, e1011327 (2023).

90. Pedregosa, F. et al. Scikit-learn: machine learning in Python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).
91. Walther, A. et al. Reliability of dissimilarity measures for multi-voxel pattern analysis. *Neuroimage* **137**, 188–200 (2016).
92. Nili, H. et al. A toolbox for representational similarity analysis. *PLoS Comput. Biol.* **10**, e1003553 (2014).
93. Kriegeskorte, N., Mur, M. & Bandettini, P. A. Representational similarity analysis—connecting the branches of systems neuroscience. *Front. Syst. Neurosci.* <https://doi.org/10.3389/neuro.06.004.2008> (2008).
94. Johnston, J. Semi-orthogonal subspaces for value mediate a binding and generalization tradeoff. figshare <https://doi.org/10.6084/m9.figshare.26065600.v1> (2024).
95. Johnston, J. et al. wj2/subspace_binding: Version corresponding to paper (paper). *Zenodo* <https://doi.org/10.5281/zenodo.12194146> (2024).

Acknowledgements

We thank B. Vinje, who led an excellent summer journal club on the binding problem at UC Berkeley in 2001. We thank S. Fusi for useful discussions of previous versions of this paper. We thank M. Wang, T. Cash-Padgett, M. Mancarella, C. Strait, T. Blanchard and B. Sleezer for assistance with data collection. We also thank L. Mickiewicz, A. Ong and A. Silcott for administrative support. This research was supported by NIDA R01 DA038615 (to B.Y.H.) and MH124687 (to B.Y.H.) W.J.J. was supported by NSF 1707398, Simons Foundation 542983SPI, Gatsby Charitable Foundation GAT3708, NIMH R01 MH129031 and the Kavli Foundation. We acknowledge computing resources from Columbia University's Shared Research Computing Facility project, which is supported by NIH Research Facility Improvement Grant

1G2ORR030893-01, and associated funds from the New York State Empire State Development, Division of Science Technology and Innovation (NYSTAR) Contract C090171, both awarded 15 April 2010.

Author contributions

W.J.J., J.M.F., S.B.M.Y., R.B.E. and B.Y.H. conceived of the project. S.B.M.Y. and B.Y.H. designed the experiments. S.B.M.Y. and R.B.E. collected the data. W.J.J. and J.M.F. designed and performed the analyses. W.J.J. developed the theoretical approach. W.J.J. and J.M.F. created the figures. W.J.J., J.M.F. and B.Y.H. wrote and edited the paper.

Competing interests

The authors declare no competing interests.

Additional information

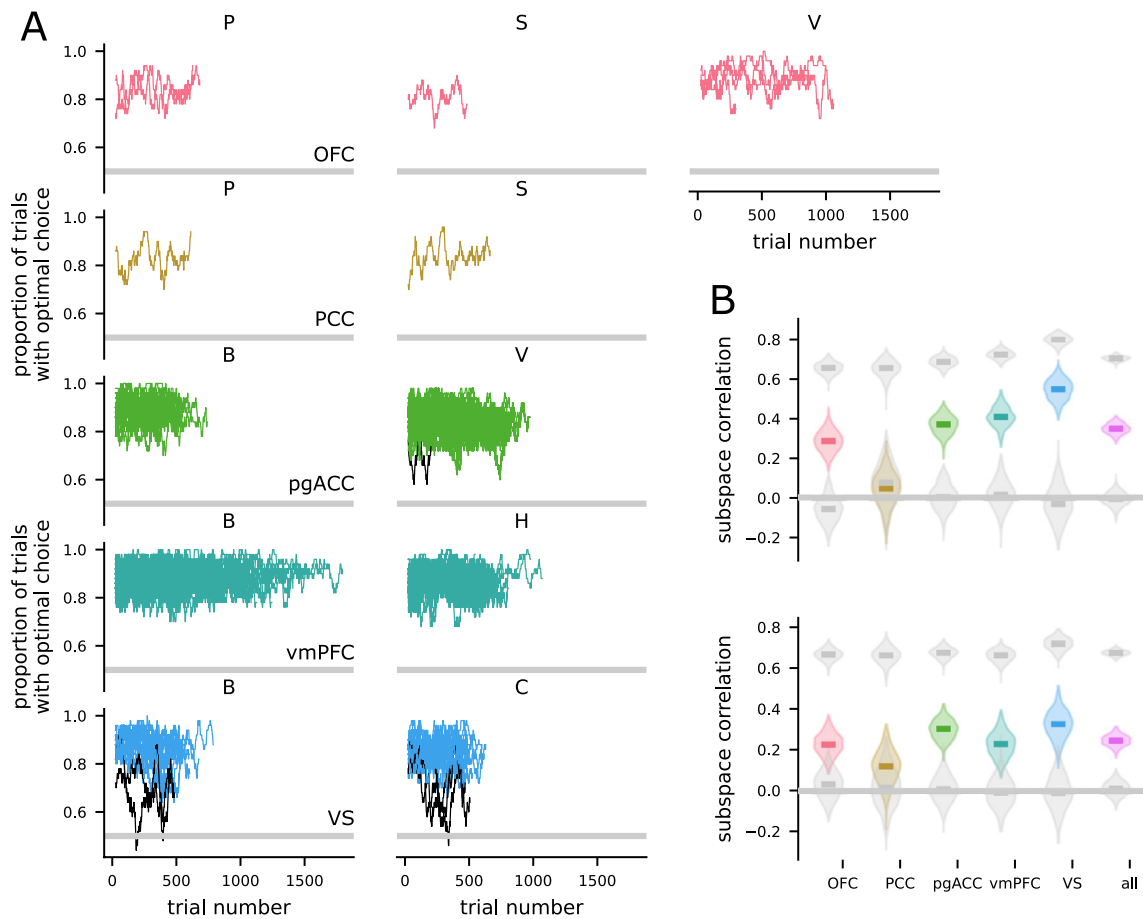
Extended data is available for this paper at <https://doi.org/10.1038/s41593-024-01758-5>.

Supplementary information The online version contains supplementary material available at <https://doi.org/10.1038/s41593-024-01758-5>.

Correspondence and requests for materials should be addressed to W. Jeffrey Johnston.

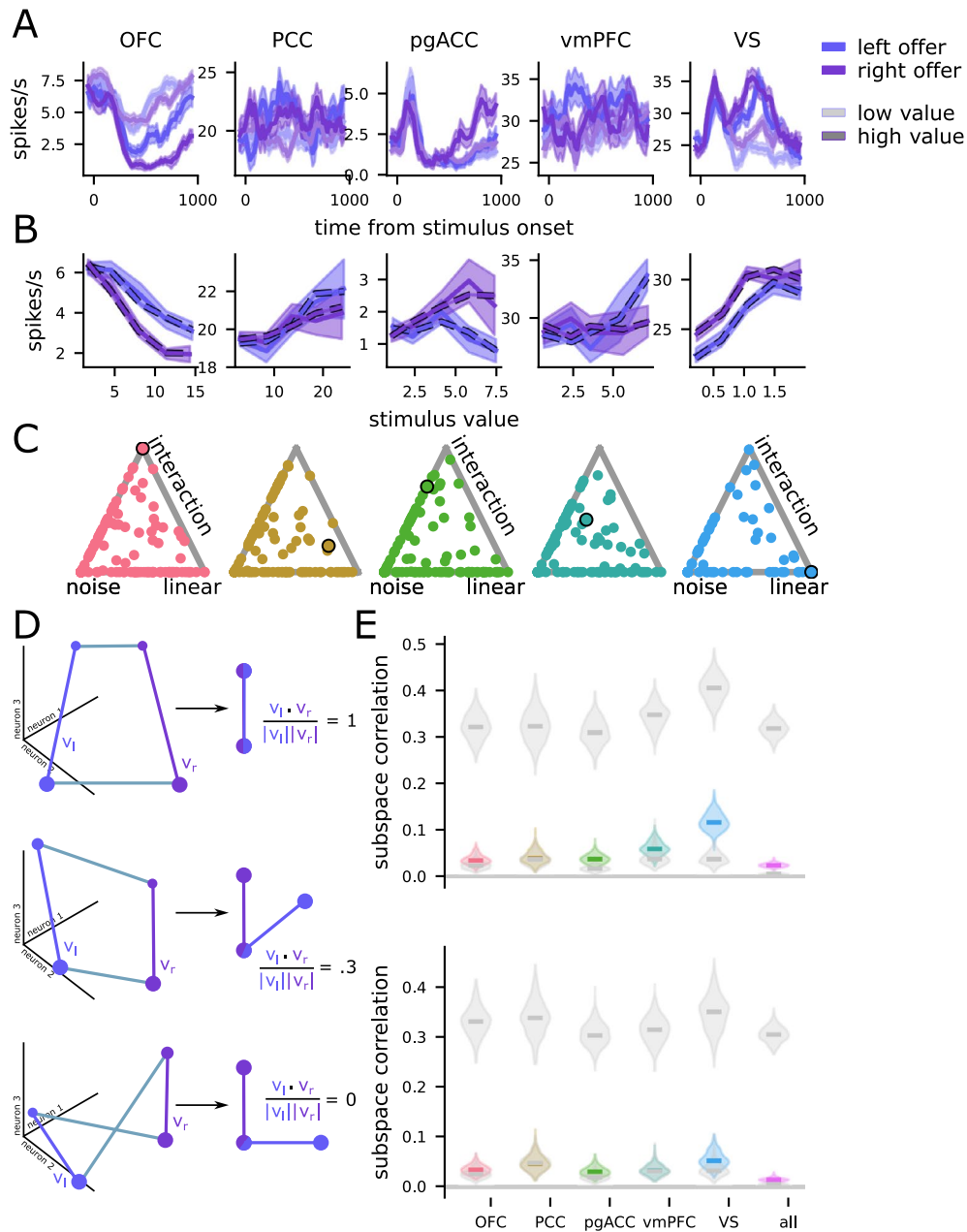
Peer review information *Nature Neuroscience* thanks the anonymous reviewers for their contribution to the peer review of this work.

Reprints and permissions information is available at www.nature.com/reprints.



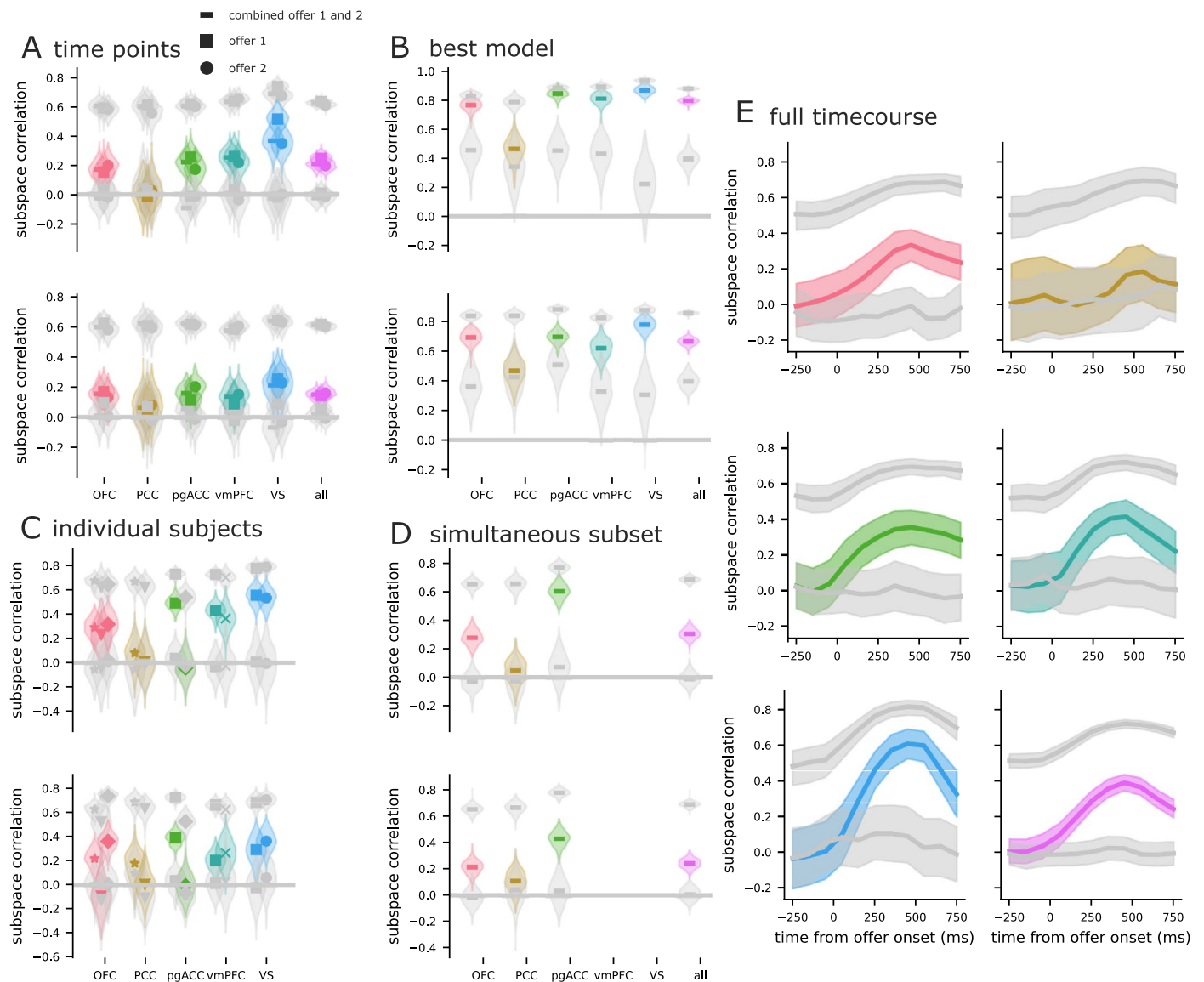
Extended Data Fig. 1 | Behavioral consistency across subjects and regions.
A. The time-average proportion of optimal choices made by each subject (columns) during experiments in each region (rows). The proportion is smoothed with a 50 trial-wide boxcar filter. The black traces ($n = 8$ across all

regions and animals) are sessions where performance dropped below a threshold of 60% optimal choices for at least one time bin. **B.** The subspace correlation analysis performed for the subset of sessions where performance did not drop below the 60% threshold. The results are similar to the full set of sessions.



Extended Data Fig. 2 | Example neurons, model comparison, and subspace correlations for the linear models with a nonlinear value representation.
A. The firing rates of example neurons from each region during the offer window, shown for high and low value offers presented on the left or right side (100 ms boxcar filter, shaded area is SEM). **B.** The value-response function for each neuron in **A**. The value-response function fit by the linear regression model with a b-spline value encoding and an interaction term is overlaid (dashed lines). The b-spline representation uses 4 knots and is degree 2. **C.** A simplex showing the weight given to each of the noise-only, linear, and interaction regression models by the Bayesian model stacking analysis. The points corresponding to the example neurons shown in **A** and **B** have dark outlines here. Both the linear and interaction categories include both linear and spline value representation models. **D.** Schematic of three different representational geometries that would

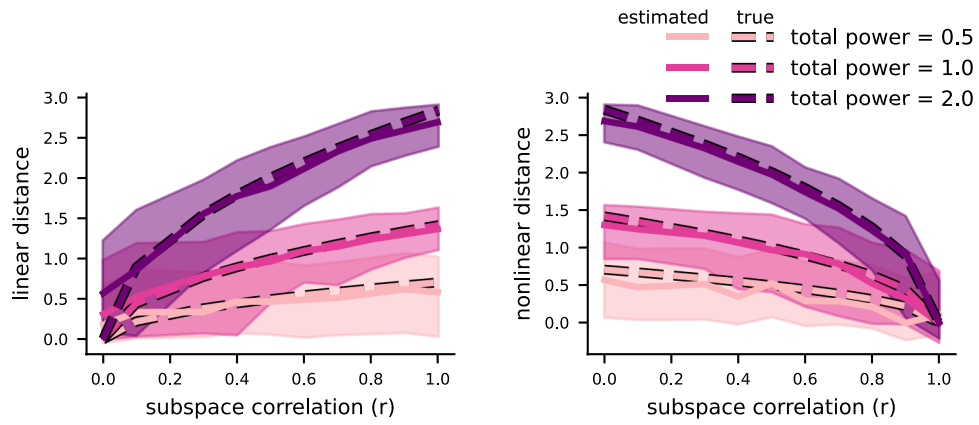
lead to different subspace correlation results. (top) Two perfectly aligned value vectors v_l and v_r in population space (left) would produce a subspace correlation close to 1 (right). (middle) Two partially aligned value vectors v_l and v_r in would produce a subspace correlation between 0 and 1 (note there is an additional possibility: partially aligned but negatively correlated subspaces; not schematized). (bottom) Two unaligned value vectors v_l and v_r would produce a subspace correlation close to 0. **E.** Alignment indices for all regions for the offer presentation window (top) and the delay period (bottom). The upper gray point is the alignment index expected if the left- and right value representations were aligned and corrupted only due to noise – the lower gray point is the noise floor. In these models, pgACC, VS, and the combined population are semi-orthogonal, while PCC, OFC, and vmPFC are indistinguishable from orthogonal.



Extended Data Fig. 3 | Variations on the subspace correlation analysis.

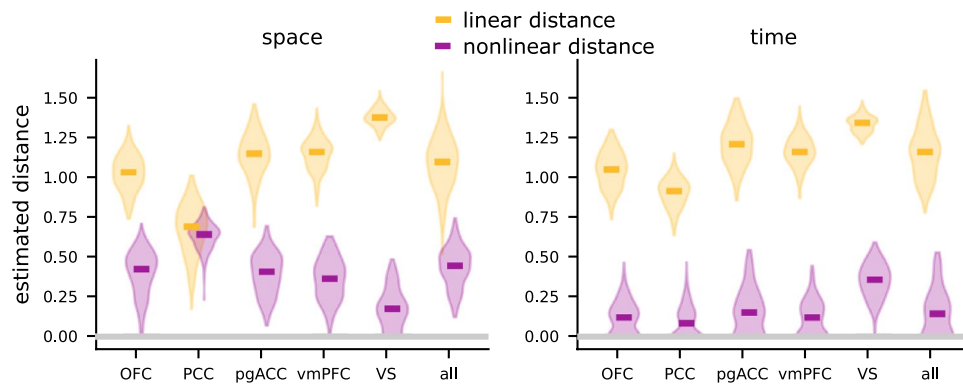
A. Subspace correlation for the different time points of the experiment (that is, offer 1 and offer 2) as well as a combined set of trials from both offers, but given the same total number of trials as either offer individually. **B.** Subspace correlation for the best fitting models given the model comparison analysis. OFC, vmPFC, VS, and the combined population (“all”) have semi-orthogonal representation, while PCC has orthogonal representations and pgACC has

parallel representations in the first time period, but orthogonal representations in the second time period. **C.** Subspace correlation for separate monkeys and regions. **D.** Subspace correlation for the set of sessions included in the simultaneous population error analysis in Fig. 5. **E.** The average subspace correlation for the full set of trials shown as a timecourse analysis; the shaded area represents the 95% confidence interval.

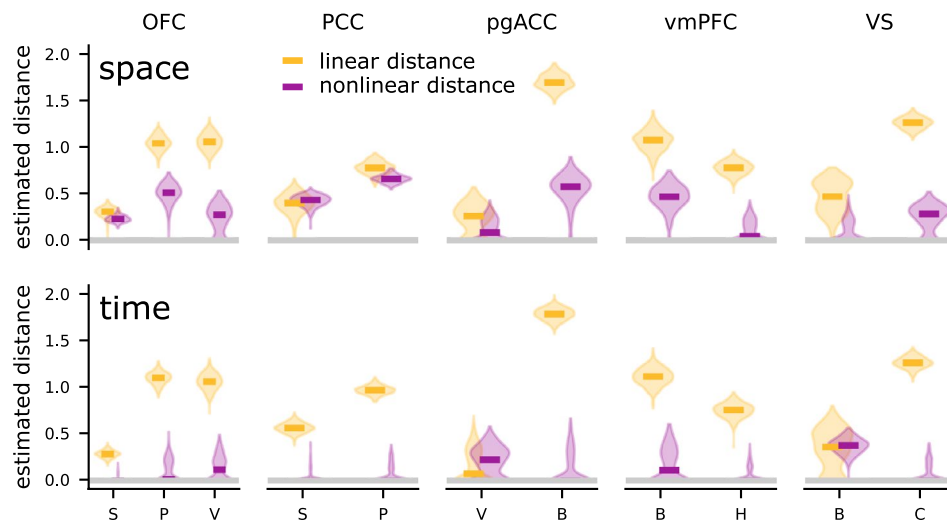


Extended Data Fig. 4 | Accurate recovery of linear and nonlinear distances from simulated data. The true (dashed line) and estimated (solid line with error bars) linear (left) and nonlinear (right) distances from simulated data. The error bars represent 95% intervals around the estimated values from $n = 100$ different

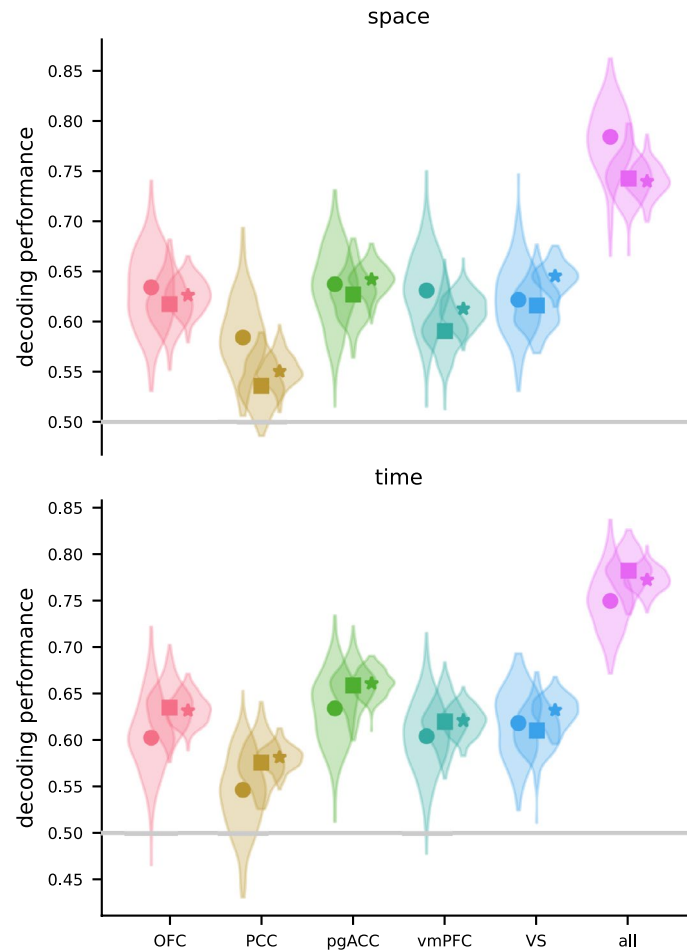
synthetic datasets with matched statistics. Our decomposition accurately recovers the linear and nonlinear distances, as the true value is always within error bars of the estimated value and the estimate is typically unbiased.



Extended Data Fig. 5 | Distance estimated from subsampled neural populations. The distances are estimated from subsampled populations of 80 neurons. Otherwise, this plot is the same as Fig. 4c.



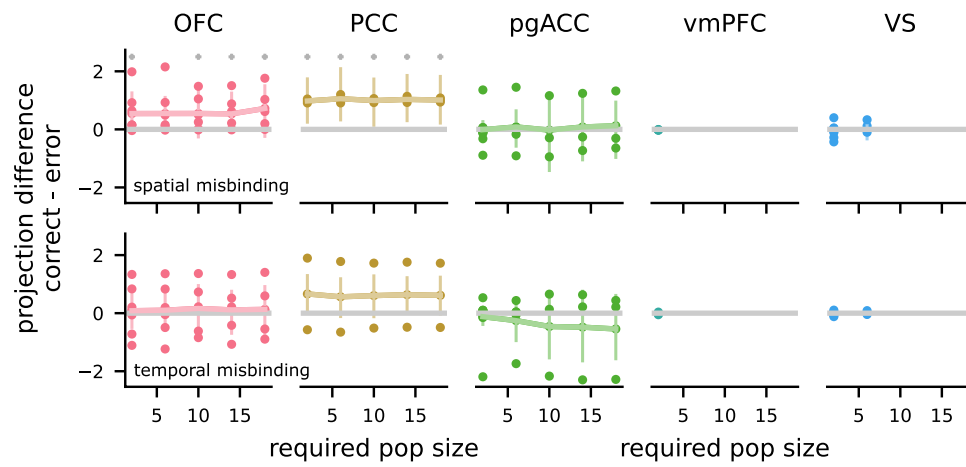
Extended Data Fig. 6 | Estimated distances for individual subjects. This figure is analogous to Fig. 4c, but where the data for each region is separated into the constituent subjects.



Extended Data Fig. 7 | Value decoding, value generalization, and predicted value generalization of the code within each recorded region. (Top)

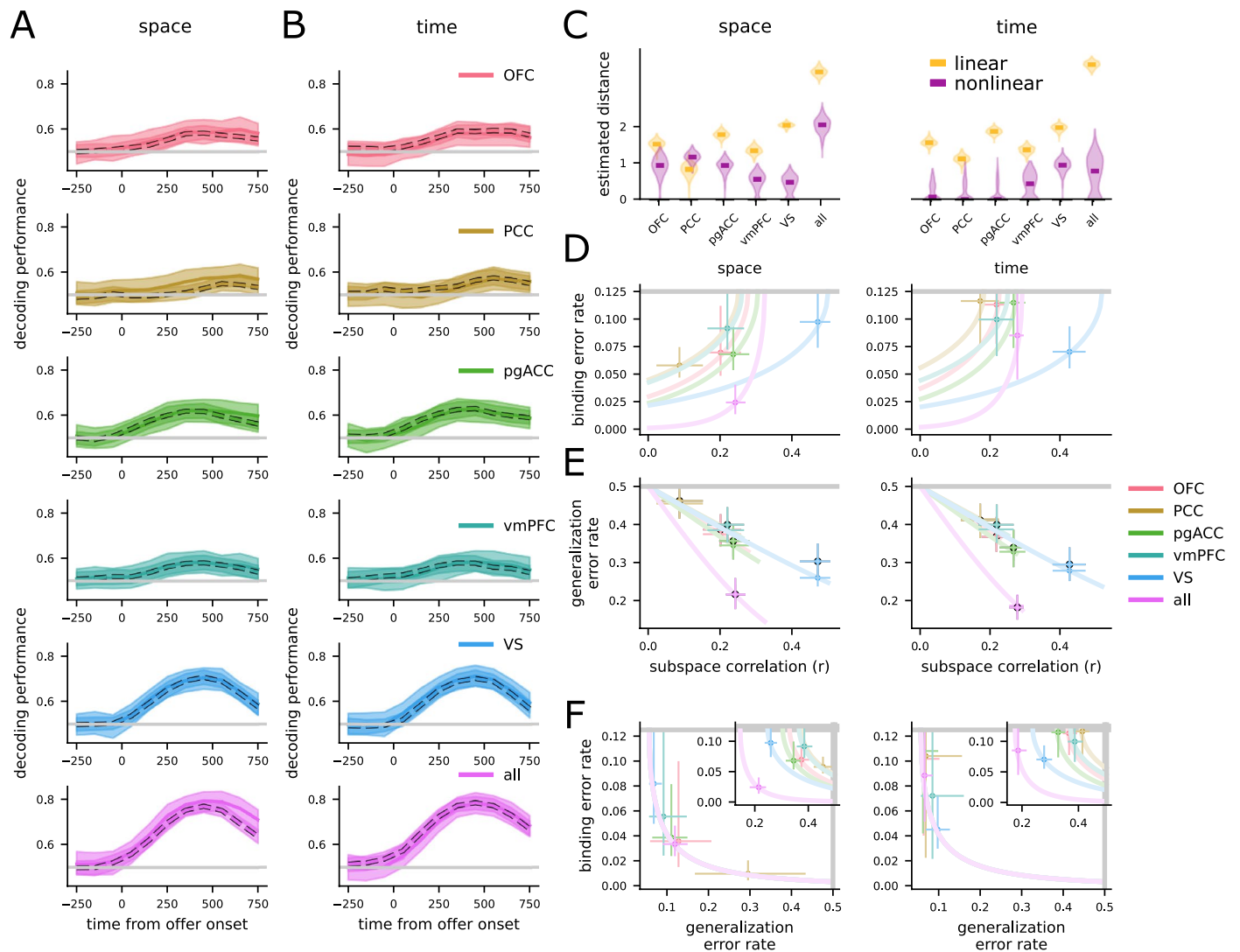
Pseudopopulation value decoding performance (circles), generalization performance (squares, trained on offers from one side, tested on offers from the other side), and predicted generalization performance (stars) shown for each

region and the neural population combined across regions (“all”), shown for the left and right value comparison. The violin plot shows the values produced from two hundred bootstrap resamples of the trials. **(Bottom)** The same as **(top)** except shown for the offer 1 and offer 2 comparison.



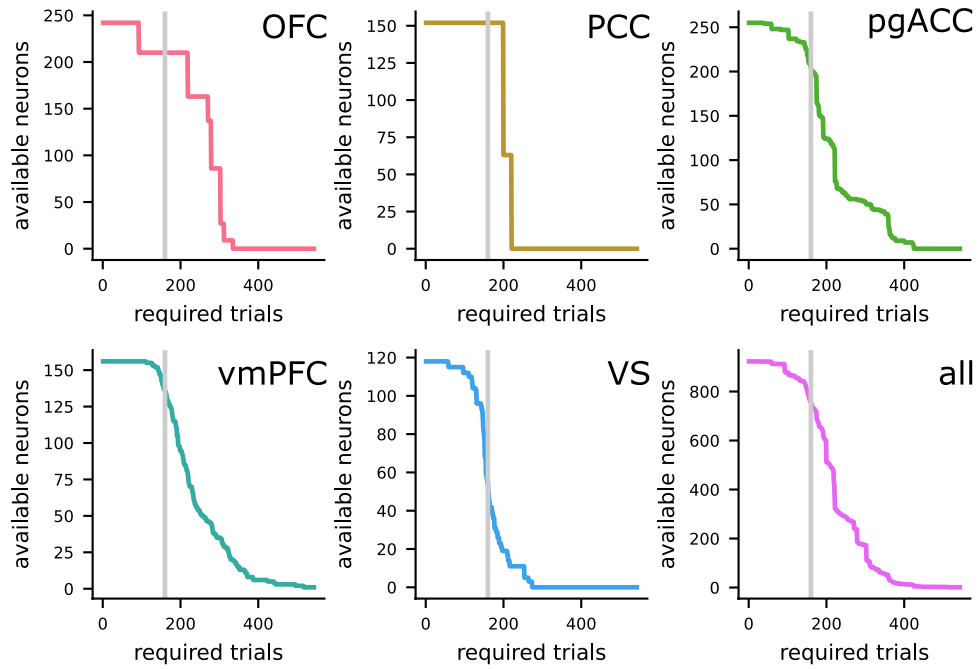
Extended Data Fig. 8 | Dependence of the behavioral decoding result (Fig. 6f) on the required number of neurons in the simultaneously recorded population. The gray crosses indicate significance under a one-sided T-test. The

individual circles show sessions, the error bar shows the average across cross-validation runs of those sessions. The gray stars indicate significance at the $p < .05$ level.



Extended Data Fig. 9 | The main results replicate when safe trials are included. This figure replicates Fig. 4 from the main text but including the safe trials in the dataset. The results are qualitatively similar in both conditions. **A**, Shows offer value decoding and generalization across left and right offers. **B**, Shows offer value decoding and generalization across offer 1 and offer 2. **C**, Shows the linear and nonlinear distances estimated for both left and right

offers (left) and offer 1 and 2 (right). **D**, Shows the predicted binding error rate for each region in both the spatial (left) and temporal (right) configurations as a function of subspace correlation. **E**, The same as (**D**) but shows the predicted and empirical generalization error rates. **F**, Shows the position of each region on the binding and generalization error rate plane.



Extended Data Fig. 10 | The available neurons for a particular required number of trials. As more trials for each condition are required, fewer neurons are available for inclusion in the pseudopopulation. The default throughout the paper is 160, shown in grey.

Reporting Summary

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

Statistics

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

n/a | Confirmed

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

Our web collection on [statistics for biologists](#) contains articles on many of the points above.

Software and code

Policy information about [availability of computer code](#)

Data collection | Psychtoolbox 3.0 running in MatLab 2012 was used for experiment delivery
Brainsight was used to confirm recording locations

Data analysis | This work relies on open source scientific computing software:

```
python 3.8.2
numpy 1.23.5
scipy 1.10.1
sklearn 1.3.0
rsatoolbox
matplotlib 3.7.2
```

as well as custom code built on these packages; the code is available on github: https://github.com/wj2/subspace_binding

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

Data

Policy information about [availability of data](#)

All manuscripts must include a [data availability statement](#). This statement should provide the following information, where applicable:

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

The data is available on figshare: <https://doi.org/10.6084/m9.figshare.26065600> under a CC BY 4.0 license

Research involving human participants, their data, or biological material

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

Reporting on sex and gender	N/A
Reporting on race, ethnicity, or other socially relevant groupings	N/A
Population characteristics	N/A
Recruitment	N/A
Ethics oversight	N/A

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

Field-specific reporting

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

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

For a reference copy of the document with all sections, see [nature.com/documents/nr-reporting-summary-flat.pdf](https://www.nature.com/documents/nr-reporting-summary-flat.pdf)

Life sciences study design

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

Sample size	six Rhesus macaques were used in these experiments; we made recordings from 5 distinct brain regions, with two different animals recorded per brain region. Recording from two different animals per region allowed us to replicate our results twice for each region, which follows the conventions of the field (e.g., Yoo & Hayden, 2020; Strait et al., 2014)
Data exclusions	Trials were excluded when the animal did not successfully complete the trial. Neurons were excluded when there were less than 160 trials per condition in the decoding and distance analyses in Fig 4 and 5. Neurons were excluded when there were less than 25 trials per condition in the distance analyses in Fig 6. Recordings sessions were excluded for the error decoding analysis if they did not include 10 or more neurons. Our results do not depend on the exact value of these exclusion thresholds, and this is illustrated in the supplement in some cases.
Replication	All results were robust to rerunning with different random seeds, and random subselections of trials as described in the text. We also tested several different choices for the exclusion criteria listed above and found qualitatively similar results in the different cases. For the experiments, our results were replicated in at least two animals for each region. The results across the two monkeys for each region were qualitatively similar and are shown in the supplement, with one exception (as discussed in the main text): the results in vmPFC appear to primarily be driven by one of the two monkeys.
Randomization	The condition for each trial was randomly chosen, which indicates that our findings should not depend on systematic instability of the recordings (which would be a concern if, for example, all trials from a certain condition were presented first). There was not randomization at the level of animals since our design is within subjects. The particular region(s) recorded from for each animal was not systematically randomized, since these data were collected over the course of several distinct projects and over several years within the lab.
Blinding	No blinding was performed, because the study design was within subjects and because it is impractical to blind investigators from the brain region they are recording from.

Reporting for specific materials, systems and methods

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

Materials & experimental systems

n/a	Involvement
<input checked="" type="checkbox"/>	<input type="checkbox"/> Antibodies
<input checked="" type="checkbox"/>	<input type="checkbox"/> Eukaryotic cell lines
<input checked="" type="checkbox"/>	<input type="checkbox"/> Palaeontology and archaeology
<input type="checkbox"/>	<input checked="" type="checkbox"/> Animals and other organisms
<input checked="" type="checkbox"/>	<input type="checkbox"/> Clinical data
<input checked="" type="checkbox"/>	<input type="checkbox"/> Dual use research of concern
<input checked="" type="checkbox"/>	<input type="checkbox"/> Plants

Methods

n/a	Involvement
<input checked="" type="checkbox"/>	<input type="checkbox"/> ChIP-seq
<input checked="" type="checkbox"/>	<input type="checkbox"/> Flow cytometry
<input checked="" type="checkbox"/>	<input type="checkbox"/> MRI-based neuroimaging

Animals and other research organisms

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

Laboratory animals	Six Rhesus macaques (<i>Macaca mulatta</i>) were used in these experiments. They were all male and aged 3 - 9.
Wild animals	No wild animals were used.
Reporting on sex	Our findings only apply to male monkeys.
Field-collected samples	No samples were collected from the field.
Ethics oversight	The surgical and experimental procedures were approved by either the University Committee on Animal Resources at the University of Rochester or the IACUC at the University of Minnesota

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

Plants

Seed stocks	<i>Report on the source of all seed stocks or other plant material used. If applicable, state the seed stock centre and catalogue number. If plant specimens were collected from the field, describe the collection location, date and sampling procedures.</i>
Novel plant genotypes	<i>Describe the methods by which all novel plant genotypes were produced. This includes those generated by transgenic approaches, gene editing, chemical/radiation-based mutagenesis and hybridization. For transgenic lines, describe the transformation method, the number of independent lines analyzed and the generation upon which experiments were performed. For gene-edited lines, describe the editor used, the endogenous sequence targeted for editing, the targeting guide RNA sequence (if applicable) and how the editor was applied.</i>
Authentication	<i>Describe any authentication procedures for each seed stock used or novel genotype generated. Describe any experiments used to assess the effect of a mutation and, where applicable, how potential secondary effects (e.g. second site T-DNA insertions, mosaicism, off-target gene editing) were examined.</i>