JNeurosci
THE JOURNAL OF NEUROSCIENCE

*Research Articles: Systems/Circuits*

# Abstract value encoding in neural populations but not single neurons

https://doi.org/10.1523/JNEUROSCI.1954-22.2023

**Cite as:** J. Neurosci 2023; 10.1523/JNEUROSCI.1954-22.2023

Received: 18 October 2022
Revised: 7 May 2023
Accepted: 11 May 2023

# **Title**: Abstract value encoding in neural populations but not single neurons

**Abbreviated title**: Transformable subspaces in economic choice

**Authors:** Justin M. Fine, David J-N. Maisson, Seng Bum Michael Yoo, Tyler V. Cash-Padgett, Maya Zhe Wang, Jan Zimmermann, and Benjamin Y. Hayden

**Affiliation**

Department of Neuroscience and Center for Magnetic Resonance Research
University of Minnesota, Minneapolis MN 55455

**Corresponding author**

Benjamin Yost Hayden
Email: benhayden@gmail.com
Current address: Department of Neurosurgery
Baylor College of Medicine
Houston, TX, 77040

**46 pages, 8 figures, 0 tables**
**Abstract:** 193 words
**Introduction:** 689 words
**Discussion:** 1392 words

**Conflict of interest:** None

**Acknowledgements**

We thank Geoffrey Schoenbaum for helpful discussions. We thank Marc Mancarella, Caleb Strait and Tommy Blanchard for assistance with data collection, Sarah Heilbronner for help with anatomy, and the rest of the Hayden and Zimmermann labs for valuable discussions. Funding provided by National Institute on Drug Abuse grant P30 DA048742-01A1 (to BYH and JZ), National Institute for Biomedical Imaging Grant P41 EB027061 (to BYH and JZ), and the UMN AIRP award (to BYH and JZ).

## ABSTRACT

An important open question in neuroeconomics is how the brain represents the value of offers in a way that is both abstract (allowing for comparison) and concrete (preserving the details of the factors that influence value). Here we examine neuronal responses to risky and safe options in five brain regions that putatively encode value in male macaques. Surprisingly, we find no detectable overlap in the neural codes used for risky and safe options, even when the options have identical subjective values (as revealed by preference) in any of the regions. Indeed, responses are not just uncorrelated but occupy distinct (semi-orthogonal) encoding subspaces. Notably, however, these subspaces are linked through a linear transform of their constituent encodings, a property that allows for comparison of dissimilar option types. This encoding scheme allows these regions to have their cake and eat it too: they can encode the detailed factors that influence offer value (here, risky and safety) but also directly compare dissimilar offer types. Together these results suggest a neuronal basis for the qualitatively different psychological properties of risky and safe options and highlight the power of population geometry to resolve outstanding problems in neural coding.

## SIGNIFICANCE STATEMENT

To make economic choices, we must have some mechanism for comparing dissimilar offers. We propose that the brain uses distinct neural codes for risky and safe offers, but that these codes are linearly transformable. This encoding scheme has the twin advantages of allowing for comparison across offer types while preserving information about offer type, which in turn allows for flexibility in changing circumstances. We show that responses to risky and safe offers exhibit these predicted properties in five different reward-sensitive regions. Together, these results highlight the power of population coding principles for solving representation problems in economic choice.

## INTRODUCTION

We are often faced with the need to choose between options that differ qualitatively. To make such choices, it is helpful to have access to an abstract representation of the value of each option. Much of neuroeconomics is predicated on the assumption that such representations must exist and, in particular, that they exist in the form of specialized *abstract value neurons* (Platt and Glimcher, 1999; Padoa-Schioppa and Assad, 2006; Kennerley et al., 2009; Lau and Glimcher, 2008; Peters and Buchel, 2010; Kolling et al., 2016). Such neurons would, by definition, have firing rates that covary monotonically with the values of offers regardless of their other qualities. That is, two options with identical subjective values would elicit the same firing rate response in an abstract value neuron even if they differed in other ways. For example, if asked to evaluate a Ferris wheel ride and an equally valued cupcake, a value neuron will necessarily have the same response to both. One advantage of the abstract neuron representational system is that a downstream decoder can produce good choices simply by identifying the neuron or neurons that encode the highest value and selecting the appropriate choice. However, while such an encoding scheme has its advantages, it has a major weakness – it is inflexible in situations where the relative importance of these features changes.

More broadly, the abstract value neuron idea is out of step with modern thinking on how information is coded in populations of neurons. While the "neuron doctrine" is focused on the idea that information explicitly encoded in the firing rates of single neurons, the "population doctrine" emphasizes the flexible and expressive power of neuronal populations (Saxena and Cunningham, 2019; Ebitz and Hayden, 2021). From the perspective of populations, abstraction can come from geometry; specifically, from transformable subspaces (Elsayed et al., 2016; Tang et al., 2020; Yoo and Hayden, 2020; Libby and Buschman, 2021). This way of generating

abstraction has the major advantage of allowing a population of neurons to simultaneously carry detailed about the features of the offers and to have an abstract representation of value.

We hypothesized that neurons in core value regions use the population approach to representing value. To test this hypothesis, we studied encoding of risky and safe options in a two-option risky choice task. Psychologists have established that our minds treat risky and safe options differently. We have a strong preference for safe options (Holt and Laury, 2002; Kacelnik and Bateson, 2013; Heilbronner, 2017), risky options differentially activate motivational and emotional factors (Loewenstein et al., 2001; Slovic et al., 2004 and 2007; Lerner et al., 2015), and risk differs qualitatively from certainty among dimensions like the prospect of learning and satisfaction of curiosity (Binde, 2013; Heilbronner and Hayden, 2013; Wang and Hayden, 2021). Population geometric encoding would allow for value comparison while also maintaining information that allows risk and surety to be treated differently. It is well established that neural responses in core value regions scale with the probability (and thus the subjective value) of risky offers (Kennerley et al., 2009; So and Stuphorn, 2010; Raghuraman and Padoa-Schioppa, 2014; McCoy and Platt, 2005; Strait et al., 2014; Azab and Hayden, 2017). Perhaps surprisingly, little of this research has addressed the critical question of whether *risky* and *safe* offers use the same coding scale, and none of it has used a dense sampling of probability space needed to retrospectively identify equally valued risky and safe offers and test how their codes are related.

We examined responses of neurons in five core reward areas in a gambling task (**Figure 1A**). We found that responses of single neurons to safe offers are not just distinct but are unrelated to responses evoked by risky offers. Instead, risky and safe offers are encoded by overlapping sets of neurons using distinct and semi-orthogonal subspaces. At the same time,

these subspaces are mutually transformable. These results are consistent with the hypothesis that the brain's core value regions strategically use subspace orthogonalization to flexibly partition different offer types in a way that allows for value comparison. More generally, these results endorse the utility of population perspectives in tackling classic problems in neuroscience.

## MATERIALS AND METHODS

*Surgical procedures.* Subjects were male macaques. All procedures were approved by either the University Committee on Animal Resources at the University of Rochester or the IACUC 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 of the animals were handled according to approved institutional animal care and use committee (IACUC) protocols (#2005-619 38127A) of the University of Minnesota. The protocol was approved by the Committee on the Ethics of Animal Experiments of the University of Minnesota (NIH permit number: A3456-01). All surgery was performed under controlled anesthesia. Six male rhesus macaques (*Macaca mulatta*) served as subjects. A small prosthesis for head fixation was used. Subjects were habituated to laboratory conditions and then trained to perform oculomotor tasks for liquid rewards. We place a Cilux recording chamber (Crist Instruments) over the area of interest (see Behavioral Tasks for breakdown). 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 we 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 atlas (Paxinos et al., 2000). In all cases we sampled evenly across the regions. Neuronal recordings in OFC were collected from *subjects P and S* (Yoo and Hayden, 2020); recordings in vmPFC were collected from *subjects B* and *H* (Strait et al., 2014); recordings in pgACC were collected from *subject B and V* (Maisson et al., 2021);

recordings from PCC were collected from *subject P and S* (Wang et al., 2021); and recording in VS were collected from *subject B and C* (Strait et al., 2016). For details, see **Figure 1B**:

We defined **OFC 13** as lying within the coronal planes situated between 28.65 and 34.05 mm rostral to the interaural plane, the horizontal planes situated between 3 and 6.5 mm from the brain's ventral surface, and the sagittal planes between 5 and 14 mm from the medial wall. The coordinates correspond to area 13m in Paxinos et al. (2000).

We defined **vmPFC 14** as lying within the coronal planes situated between 29 and 44 mm rostral to the interaural plane, the horizontal planes situated between 0 and 9 mm from the brain's ventral surface, and the sagittal planes between 0 and 8 mm from the medial wall. These coordinates correspond to area 14m in Paxinos et al. (2000).

We defined **pgACC 32** as lying with the coronal planes situated between 30.90 and 40.10 mm rostral to the interaural plane, the horizontal planes situated between 7.30 and 15.50 mm from the brain's dorsal surface, and the sagittal planes between 0 and 4.5 mm from the medial wall (**Figure 1B**). Our recordings were made from central regions within these zones, which correspond to area 32 in Paxinos et al. (2000).

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 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. (2000).

We defined **VS** as lying within the coronal planes situated between 20.66 and 28.02 mm rostral to the interaural plane, the horizontal planes situated between 0 and 8.01 mm from the ventral surface of the striatum, and the sagittal planes between 0 and 8.69 mm from the medial wall. Note that our recording sites were targeted towards the nucleus accumbens core region of

the VS.

We confirmed recording location before each recording session using our Brainsight system with structural magnetic resonance images taken before the experiment (0.5 voxels). 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 with an error of ~1 mm in the horizontal plane and ~2 mm in the z-direction.

*Electrophysiological techniques.* Either single (FHC) or multi-contact electrodes (V-Probe, Plexon) were lowered using a microdrive (NAN Instruments). Individual action potentials were isolated on a Plexon (Dallas, TX) or Ripple (Salt Lake City, UT). Neurons were selected for study solely based on 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 (Mathworks) with Psychtoolbox and Eyelink Toolbox. Visual stimuli were colored rectangles on a computer monitor placed 57 cm from the animal and centered on its eyes (Fig. 1*A*). A standard solenoid valve controlled the duration of juice delivery. Solenoid calibration was performed daily.

*Behavioral tasks.* Six macaques performed in the risky choice task. Both tasks 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 (Hayden et al., 2010; Blanchard et al., 2014).

*Risky choice task.* 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; Fig. 2*A*). Two parameters defined gamble offers, *stakes* and *probability*. Each gamble rectangle was divided into two portions, one red and the other either gray, blue, or green. The size of the color portions signified the probability of winning a small (125 μL, gray), medium (165 μL, blue), or large reward (240 μL, 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-stakes) gamble, a 43.75% probability of green (high-stakes) gambles, and a 12.5% probability of gray options (safe offers).

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 subject 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 subjects 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 sec 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.

*Estimation of subjective value equivalence*. We calculated the indifference point between

safe and risky offers. For each subject, independently, we fitted a sigmoidal function to the distribution of choices (safe or risky) across the full range of risky offer probabilities.

$$f(x) = \frac{a}{b + e^{bx}}$$

where x = the probability associated with the risky offer, and *f*(x) = the likelihood of choosing the safe offer; a (the maximum value of the curve) and b (the growth rate; steepness) are coefficients of the function, estimated by the fitting procedure for maximizing $R^2$. Using the fitted sigmoidal function, we then estimated the value of x needed to produce a safe choice likelihood of 0.5; that is, a risky offer choice is equally likely as a safe choice. This is called the *indifference point*. We performed this analysis separately for medium- and high-stakes gambles, and separately for each subject. Note that these curves are the result of a fitting procedure that has greater error in fitting the asymptotes. This is due to the small number of variant trials at these asymptotes that the algorithm is highly sensitive to outliers at these values.

*Statistical methods.* We constructed peristimulus time histograms by aligning spike rasters to the presentation of the first offer and averaging firing rates across multiple trials. We calculated firing rates in 20 ms bins, but we analyzed them in longer (500 ms) epochs. Some statistical tests of neuronal activity were only appropriate when applied to single neurons because of variations in response properties across the population. We have used this epoch in all our past research on this and similar tasks (Strait et al., 2014, 2015, 2016; Azab and Hayden, 2017, 2018, 2020). We have found that this epoch provides a good characterization of functional responses and allows for fair comparison across brain regions. We used it here for those reasons and because adherence to a single pre-planned epoch of interest reduces the likelihood of inadvertent "p-hacking".

*Shuffle analysis*: to understand the values of correlations that we would obtain by chance,

in the absence of a true effect, is it useful to repeat our analysis on dummy data. To do so, we performed a specific control procedure developed by Elsayed et al. (2016). In this analysis, we wished to preserve the features of each neuron without removing that information, so we shuffled across time, meaning that we randomly sorted the spikes to occur at random times, but maintained their linkage to the neuron that generated them. This procedure then gave dummy data that we performed our correlation analyses on. More specifically, we shuffled the data from across all three of the previously computed response matrices (safe, matched-medium, and matched-large) and computed the alignment index between pairs of shuffled sets. Because the data are randomized, any alignment index at or below this dummy threshold would be considered at least semi-orthogonal. We then, following standard logic of bootstrap analyses, repeated this process over 1000 randomizations. Then, to test for significance, we computed the 99% confidence interval across iterations; a value outside of this range can therefore be said to be significant at $p < 0.005$ (given that our t-test is assumed to be one-tailed).

*Population overlap.* We adapted a previously published procedure (Azab and Hayden, 2017). For each neuron, we isolated both medium- and high-stakes risky offers with subjectively equal values to the respective subject. For each trial, we computed the average firing rate across a 500 ms window, starting from 100 ms after the appearance of the first offer (see above). We then regressed the firing rates on the corresponding offer probability on each trial to produce a beta weight representing that neurons degree of linear encoding of offer probability. For safe offers, we were limited by the fact that there is only a single probability value (100%) for safe offers. Therefore, we also computed the average firing rate from a 500 ms window that preceded the start of the trial by 100 ms. For each trial, we computed the difference between the offer response and the pre-trial response, to represent the degree of linear encoding for the safe offer

type. Finally, we performed a correlation between the absolute value of the safe offer encoding and the absolute value of the risky offer beta weights. We plotted the two encoding weights against each other and computed the line of best fit, given the correlation coefficient.

Decoding analysis. We built a pseudo-population of pseudo-trials. First, we isolated neural responses to the safe offers and the equivalent risky offers. Then, we collapsed the firing rates for each trial into an average for the 500 ms offer 1 epoch. We randomly selected 1000 samples for each neuron, under both risk conditions, resulting in two n X 1000 matrices (one for each offer type; safe vs. risky), where n represents the number of neurons recorded from each region. We used an elastic-net logistic GLM to decode safe versus risky trials. Elastic-net offers regularization over the decoding weights, including both L1 and L2-norms of both Lasso and Ridge regression, respectively. Elastic-net GLMs can provide parameter estimates with reduced bias and variance when there is collinearity in the predictors (Zou and Hastie, 2003). Specifically, the Lasso regularization has the effect of shrinking weakly contributing predictors to 0. The ridge regularization has the effect of reducing the size of all predictors. The elastic-net represents a mixture of both these costs in determining GLM parameters. We fit the GLM across 1000 within-sample bootstraps. On each iteration, a randomly selected subset, constituting half of each original safe vs. risky offer type pseudo-population, was used to fit the model and the other half was used for cross-validation. The resulting cross validation provided a measure of accuracy with which the elastic-net logistic regression model was able to decode safe vs. risky offer types from the cross-validation set of firing rates.

We performed this entire procedure on 1000 bootstrapped iterations of randomly constructed pseudo-populations. We averaged the accuracy across these out-of-sample bootstrapped cross-validations and calculated the standard error across iterations. This procedure

was performed independently for safe vs. medium-risky and safe vs. high-risky, and was performed independently for each subject. Finally, the entire procedure was performed on shuffled data, to confirm that the expected prediction accuracy for a binary decoder of randomized data would be 50%. To do this, each original pseudo-population was instead comprised of a randomly selected trial, regardless of the offer type. Additionally, each pseudo-trial was randomly assigned a label as one of the possible offer types. We then compared the average decoding accuracy between safe-risky and the shuffled data, using a standard two-sample t-test, with Bonferoni corrections.

*Subspace alignment*. We followed the procedure described in Elsayed et al., 2016. Specifically, for each structure, we separated offers and neural responses by their risk profile (safe and equivalent risky offer of medium- and high-stakes), as described previously. For each neuron, we identified two factors to incorporate into a single condition: time and offer position. Time included the same 500 ms period following the onset of offer 1 and preceding the onset of offer 2. Time was segmented into 20 ms bins. For each 20 ms bin, we computed the mean firing rate across trials on which the offer was positioned on either the left or right of the screen. Thus, we constructed a condition (time X offer position) X neuron matrix of mean firing rates; that is, a 50 X n-neurons matrix. One such matrix was constructed for safe offers, one for medium, and one for high-stakes risky offers of equivalent value. Firing rates in prefrontal areas of macaques tend to be sparse. So, we smoothed these matrices using a gaussian filter, with a sigma equal to one. We smoothed across columns in the matrix; in other words, we smoothed each neurons individually. We then normalized the smoothed matrices, by computing the z-score within each cell, to account for differences in encoding scaling between neurons.

Next, we performed a principal component analysis, using eigenvalue decomposition, on

293 the safe response matrix, providing a transformation matrix into which we projected both the
294 safe response matrix and each of the risky response matrices. We computed the explained
295 variance due to each of the principal components. We performed the same process of
296 dimensionality reduction for each of the risky offers, projecting both the safe response and
297 corresponding risky response data into the resulting principal component spaces (medium- and
298 high-stakes risky response each into their own principal component spaces). To determine if the
299 subspaces were aligned, we computed an alignment index:
300

$$A_{idx} = \frac{Tr\left(D_{risky}^{T} C_{safe} D_{risky}\right)}{\sum_{i=1}^{sel-dim} \sigma_{safe}(i)}$$

301 where Tr() is the sum along the diagonal entry, sel-dim = the number of selected principal
302 components (or ten, in the current study), $D_{risky}$ are the set of top sel-dim eigenvectors, $C_{safe}$ is the
303 covariance matrix for the safe responses, $\sigma_{safe}(i)$ is i-th singular value of $C_{safe}$. Essentially, the
304 variance explained in safe responses by the top ten principal components of the risky responses
305 is normalized against the sum of the variance explained by the top ten principal components of
306 the safe responses. Note that we also performed this calculation using both the top 4 and top 7
307 principal components. This control did not change the results of the significance tests, and so
308 they are not reported.
309 In the Results, we described the difficulty associated with interpreting alignment indices
310 given the structure of the data, and the corresponding need to use shuffling procedures to
311 determine the effective upper and lower limits of measured alignment. To implement this
312 procedure, we tested whether the safe and risky subspaces were more or less orthogonal, relative
313 to randomly sampling within the space of this fixed correlation structure. We concatenated all
314 safe and risky offer data into a single matrix. We then computed the covariance matrix and

315 performed an eigenvalue decomposition for the covariance matrix. We then randomly sampled
316 subspaces that were aligned to the fixed correlation structure of the response space, as follows:

$$v_{align} = \frac{ortho\left(U\sqrt{Sv}\right)}{||U\sqrt{Sv}||_2}$$

317 where U and S are the eigenvectors and eigenvalue matrices, respectively, of the
318 computed covariance matrix. A matrix (v) was drawn from a normal distribution with a mean =
319 0.0 and variance = 1.0. *Orth*() computes the orthonormal basis of the projected matrix. This
320 process essentially maintains the neuronal covariance structure of the original covariance matrix
321 used for the eigenvalue decomposition. We repeated this process across 1000 iterations and
322 computed the alignment index for each, according to the above description. We then calculated
323 the average alignment index and the 99% confidence intervals across the 1000 iterations.

## RESULTS

### Behavior

On each trial of the *risky choice task*, a macaque (*Macaca mulatta*) subject chose between two offers that varied in magnitude and probability (**Figure 1A**, Strait et al., 2014, see **Methods**). Safe offers (12.5% of offers) provided a small volume of juice (125 $\mu$L) with 100% certainty. Risky offers provided either a medium (165 $\mu$L, 43.75% of offers) or large (240 $\mu$L, 43.75% of offers) volume of juice with a defined probability (0-100%, 1% increments). The offer types for the two offers were generated at random and independently on each trial. Our dataset consists of responses from six subjects in 315 recording sessions and consists of 211,884 trials (average 672.6 trials per session). Target regions for neural recordings are illustrated in **Figure 1B** and anatomical boundaries are provided in the **Methods**.

Subjects consistently performed at a high level, were modestly risk-seeking, and did not differ from each other qualitatively (**Figure 1C**). Details of species-typical behavior in this task are given elsewhere (in most detail in Farashahi et al., 2018 and 2019). Results of these analyses are not repeated here, except to confirm, as we have previously shown, that subjects' behavior is quite stable and consistent within subjects, both across and within sessions, and across subjects and sessions **(Figure 1D-F**). Indeed, all subjects showed the same behavioral patterns we have observed using this task in past studies.

All analyses in this paper make use of subjective values instead of expected values. To identify the relative values of safe offers, we computed the risky-safe indifference point. We calculated, separately for each subject and separately for medium and large stakes offers, the likelihood that the subject would choose the safe offer as a function of the probability associated with the risky offer. We then calculated indifference using a standard approach in which we fit

the resulting data with a sigmoid curve and calculated the point at which the best-fitting curve crossed the indifference line (see **Methods**).

As in all of our past studies using this task, all subjects were risk-seeking. The cross-subject average indifference point for medium-stakes risky offers corresponded to an offer probability of 0.33 ± 0.05 (standard deviation). A risk-neutral subject would have had an indifference point at 0.76. The average indifference point for high-stakes risky offers corresponded to an offer probability of 0.11 ± 0.04. A risk-neutral subject would have had an indifference point of 0.52. As we have observed many times, preferences were highly consistent across many contexts. For example, indifference points are similar for the first and second offers (offer 1: medium: 0.34; high: 0.11; offer 2: medium: 0.22; high = 0.11), for offers made early and late in the session (early: medium: 0.29; high: 0.09; late: medium: 0.31; high: 0.13), and when risky offers appear on the left or right (left: medium: 0.27, high: 0.11; right: medium: 0.36, high: 0.12). Data for an example subject are shown in **Figure 1D-F**. Data for individual subjects are shown in **Figure 2**.

**Lack of pupil size effects**

Because our central research goal is to ascertain the influence of offer type on neural activity, we wanted to ensure that our effects were not due to attention or arousal. We therefore examined the relationship between two proxies for attention, looking at time and pupil size in four of our subjects (subjects B, H, P, and S; these are the subjects for which we had pupil size recorded). We found no detectable relationship. Specifically, during the offer 1 epoch, subjects fixated the first offer for 198.5 ms if it was risky and for 199.1 ms if it was safe. These two are not different ($p=0.85$ for the group, p-values were the same or higher for all four subjects).

During the offer 2 epoch, subjects fixated the second offer for 222.4 ms if it was risky and for 220.9 ms if it was safe (again, these are not different, p=0.73). During the choice epoch, we calculated the total looking time and, again, found no differences (302.7 ms for chosen risky and 302.5 for chosen safe, p = 0.93). Nor were any of the individual subjects' fixation time differences statistically significant for any of these measures ($p > 0.05$ in all cases). These results do not appear to be dependent on our choice of epoch; we used longer analysis (1 sec) epochs for each of these analyses and found qualitatively matched results.

During the offer 1 epoch, the pupil size did not differ for risky and safe options in any of the four subjects (p=0.10 for subject S and $p > 0.5$ for the other three subjects). Specifically, relative to the baseline value (defined as 0), risky offer-evoked pupil size was -0.92 z-score units (for all subjects averaged); the corresponding value for safe offers was 0.93 z-score units. Likewise, during the offer 2 epoch, relative to the baseline value (again, defined as 0), risky offer-evoked pupil size was -0.81 z-score units (for all subjects averaged); the corresponding value for safe offers was 0.80 z-score units. These are not different for any subject (p = 0.21 for subject H, p = 0.39 for subject P, $p > 0.5$ for the other two subjects). Finally, during the choice epoch, relative to the baseline value of 0, risky choice-evoked pupil size was -0.79 z-score units (for all subjects averaged); the corresponding value for safe offers was also 0.79 z-score units. These are not different for any subject ($p > 0.5$ in all cases).

**No special subpopulations for risky-and safe-preferring neurons**

We recorded responses of 843 neurons in five brain regions while our subjects performed the risky choice task: vmPFC (area 14, 156 neurons), OFC (area 13, 157 neurons), pgACC (area 32, 255 neurons), PCC (area 29/31, 151 neurons), and VS (nucleus accumbens, 124 neurons).

We recorded in two subjects for all areas, although different subjects were used for the different areas (see **Methods**). Detailed analyses of responses to risky offers were reported previously for vmPFC and VS (Strait et al., 2014 and 2015). We have never previously examined neural responses to safe offers.

For these and subsequent analyses, we used each individual's subjective indifference point. We then defined a range of probabilities (± 2.5%, total range of 5.0%) and treated all offers within that range as being subjectively equivalent to the safe value. We subsequently checked for robustness by repeating the following analyses using a larger range (± 5%, total range of 10%) but because we found no qualitative differences, we do not report those results. We also repeated all analyses using within-session estimates of subjective value (rather than within-subject, across-session). Again, because we found no qualitative differences, we do not report those results.

Example neurons *vmPFC.70* and *pgACC.17* (**Figure 3A and B**) showed responses to safe offers whose magnitude could not be predicted from responses to risky offers. Additional sample cells, from each targeted area, including those that showed positive and negative monotonic tuning for risky values and, again, unrelated responses to safe offers (**Figure 3C-E**).

If neural responses to risky and safe offers differ, one reason may be that they make use of different subpopulations of neurons, ones specialized for risky and safe values. To determine whether the brain makes use of separate subpopulations of neurons specialized for encoding the values of safe and risky options, we adapted the approach we developed previously (Azab and Hayden, 2017 and 2018). Specifically, we have shown that if there are two categorically distinct populations of neurons, defined by distinct tuned variables (e.g., selectivity for offer 1 and selectivity for offer 2), then correlating the unsigned regression weights (betas) will produce a

negative (<0) correlation. Conversely, if the two variables are found in a single population, then correlating their unsigned betas will produce a positive correlation (**Figure 4A**). If the variables are distributed at random in the sample of neurons, then correlation will produce a null result. Here, to perform this calculation, we regressed probability for each risky option against firing; for safe options, we computed the response evoked by the presentation of the safe offer by comparing the safe-evoked response to the pre-trial baseline firing rate (see **Methods**). We found that, in the OFC, the degree to which the neural populations encode safe offers was positively correlated with the extent to which they encode matched risky offers (medium-stakes: $r = 0.49$, $p < 0.001$; high-stakes: $r = 0.29$, $p < 0.001$; Pearson's correlation; **Figure 4B**). We found similar results in all 5 other areas ($p < 0.001$, in all cases, Pearson's correlation; **Figure 4C-F**).

Note that, due to the non-normality of the data, it is possible that these correlations are unduly driven by the few outlier datapoints. We therefore repeated these correlations using Spearman's correlation. All reported correlations are still significant. Specifically, for the safe vs. medium, we find values of OFC: $\rho=0.433$; $p=0.007$; vmPFC: $\rho=0.501$; $p<0.0001$; pgACC: $\rho=0.318$; $p=0.008$; PCC: $\rho=0.498$; $p<0.001$; VS: $\rho=0.544$; $p<0.001$. For the safe vs. high, we find values of OFC: $\rho=0.288$; $p=0.01$; vmPFC: $\rho=0.401$; $p<0.008$; pgACC: $\rho=0.438$; $p=0.001$; PCC: $\rho=0.543$; $p<0.001$; VS: $\rho=0.511$; $p<0.001$. These results are consistent with more detailed studies from our lab using these and similar datasets showing no intrinsic categories in responses (Blanchard et al., 2018).

**Offer type (safe vs. risky) is decodable in all five regions**

Our principal question is how the brain might be able to disambiguate two offers that are behaviorally indistinguishable but constituted by qualitatively distinct risk features. Neurons in these regions may – and likely are - tuned for multiple variables (Rigotti et al., 2013; Raposo et al., 2014; Fusi et al., 2016; Blanchard et al., 2018; Johnston et al., 2020). This fact is important because if we assume a neuron is only tuned to a single variable, we may under-weight or ignore its other quite real tunings, which would introduce biases - likely strong ones - into our analysis. To deal with this problem, we use an elastic-net regression, which is a regularization procedure that modifies a generalized linear model (GLM) with two regularization terms: the L1 and L2 norms of both Lasso and Ridge regression, respectively. The Lasso term increases sparsity by penalizing weakly contributing neurons, while the Ridge term increases variance by reducing the size of all predictors. The elastic net GLM can provide parameter estimates with reduced bias and variance when there is collinearity in the predictors (Zhou and Hastie, 2003). We used the model as a logistic decoder to predict whether an offer was safe or risky based on the fitted model for population firing rates, using a bootstrap approach (see **Methods**).

We found that in the OFC, the elastic-net was able to decode the safe vs. medium-risky offer type from population activity with 94.9% accuracy from subject P and 90.4% accuracy in subject S (safe vs. high risky: 91.2% accuracy in subject P; 87.9% accuracy in subject S). We found similar results in the other 5 brain areas (**Figure 5**).

To determine if decoding accuracy was significant, we repeated the process after shuffling the pseudo-populations by randomly selecting from each offer type and each offer epoch, and randomly assigning a safe or risky label to the response (see **Methods**). In OFC, the decoding accuracy from shuffled data was 49.8% for both subject P and subject S, which constitutes the expected random 50/50 guess for a binary decoder. Decoding safe vs. medium-

risky offer type was significantly higher than chance (Students t-test, subject P: $t = -861.4$, $p < 0.001$; subject S: $t = -1058.8$, $p < 0.001$; even after Bonferoni correction). We found similar results in all 5 other areas, across subjects and risky offer magnitudes ($p < 0.001$ in all cases).

**Orthogonal response subspaces for value-matched risky and safe offers**

The underlying connectivity of neuron ensembles can effectively constrain activity to be correlated, rendering a low-dimensional subspace (Gallego et al., 2017; Ebitz and Hayden, 2021). Successfully decoding the safe vs. risky identity of equally valued offers could then be accomplished through their organization into distinct subspaces. We adapted previously used approaches to characterize the uniqueness of safe and matched-risky subspaces (Elsayed et al., 2016; Yoo and Hayden, 2020). Specifically, we projected risky offer responses into the safe offer response subspace and computed the explained variance for each (**Figure 6**).

When we project data from the risky subspaces (medium and high), they have lower values than when we project the safe values into the same subspace (**Figure 6A**), indicating the presence of coding orthogonality. If the brain used collinear codes for risky and safe offers, then the two sets of lines would be the same height. Indeed, a plot of the difference between the average of the two sets of times (**Figure 6B**) shows that in all cases, these are greater than zero – if the risky and safe subspaces were collinear, the differences here would be precisely equal to zero. Note that these data represent the difference between the measured value and the shuffle

alignment index; this means that data with higher shuffle alignment indices will appear lower as a result. Indeed, these results should not be taken to imply that the safe vs. high alignments are consistently greater than the safe vs. medium alignments, even though it appears this way. However, these plots are made relative to the shuffled values, which have consistent differences related to the idiosyncratic properties of the statistics of their spike trains. Instead, the only strong conclusion that can be drawn from these data is that both medium and high risky offers are consistently encoded in subspaces semi-orthogonal to those of safe offers. Note also that the projection weights are not necessarily reflective of the amount of orthogonality in a linear sense, meaning two indices that are close to each other but significantly different are not necessarily quantitatively similar.

We used these projections to quantify the extent to which subspaces were aligned ($A_{idx}$; see **Methods**). Purely orthogonal subspaces would have an alignment of 0; collinear ones would have an alignment of 1.0; semi-orthogonal subspaces would have an intermediate value between 0.0 and 1.0. Note that semi-orthogonal subspaces would be sufficient to produce separate representations and would satisfy our hypotheses. The alignment analysis indicated that in OFC, for example, comparisons of safe and medium-stakes risky response subspaces had an alignment index of $A_{idx} = 0.231$. Safe and high-stakes subspaces had an $A_{idx} = 0.266$.

These numbers do not account for the measurement limits imposed the by structure of the data. We performed two control procedures, one to determine the practical upper measurable alignment index, and the other to determine the practical lower measurable alignment index. To determine the upper limit, we performed a control procedure to measure a threshold below which indices would be considered at least semi-orthogonal (see **Methods**; Elsayed et al., 2016). Any alignment index at or. To do this, we shuffled the data from across all three response matrices

(safe, matched-medium, and matched-large) and computed the alignment index between pairs of shuffled sets (see **Methods**). We repeated this process over 1000 randomizations. Then, to test for significance, we computed the 99% confidence interval across iterations; a value outside of this range can therefore be said to be significant at $p < 0.005$ (given that our t-test is assumed to be one-tailed). We found that the average shuffled alignment index in OFC was $A_{idx} = 0.276$. Both the safe-medium and safe-high alignment indexes were below the 99% confidence interval (0.271 - 0.281) and thus both significant at $p < 0.005$ (or $p < 0.01$ with a two-tailed t-test).

To determine the lower limit, we repeated the shuffle procedure, but randomized across all axes. This analysis approach, in effect, identifies the alignment index we would observe if the data were entirely orthogonal. In OFC, the alignment index across totally shuffled data was $A_{idx} = 0.091 \pm 0.009$ (99% confidence interval). The safe-medium and safe-high alignment indexes are both quite a bit greater than this noise floor (thus both significant at $p < 0.005$). In other words, response subspaces for safe and equally valued risky offers in OFC are less orthogonal than random data, indicating that they are partially, albeit not completely, aligned. We found similar results in all structures (below the 99% confidence interval, in all cases; $p < 0.005$). We can conclude, then, that these responses have an intermediate level of alignment.

Note that it is worth reading these numbers with a good degree of caution. The space of orthogonality measures is inherently non-linear – for example, a value of 0.8 is not in any meaningful sense twice as collinear as a value of 0.4. Indeed, a quantitative interpretation of these numbers would require strong assumptions about the decoding process used by downstream structures. Nonetheless, measures of statistical significance here are interpretable and meaningful. Thus, when we say that the measured values are intermediate, we do not mean that they lie roughly halfway between orthogonal and collinear; instead, we mean that they are

neither at one extreme nor the other. Moreover, it is worth pausing to emphasize what can and cannot be inferred about orthogonality of subspaces from these numbers. On one hand the results of the significance tests can be interpreted in a conventional way – a significant difference between and upper and lower bound can be taken as evidence of semi-orthogonality. On the other hand, the magnitude of the effect comes with several caveats that make interpreting it, beyond the results of significance tests, difficult. This magnitude depends on, for example, the signal-to-noise ratio in that population of neurons, their baseline firing rate, and the particular nature of their distribution.

Finally, we asked how neural responses to medium- and high-states risky offers related to each other. If the qualitative difference in codes for risky and safe offers reflect, even in part, the qualitative differences between risk and safety, than two different risk offers should differ less than the risky and safe offers. We therefore repeated the above analyses on medium- and high-stakes risky offers. In OFC, we found that the medium- and high-stakes risky responses had an alignment index of $A_{idx} = 0.275$. This value is significantly greater than the value of the medium-risky and high-risky indices (0.231 and 0.266, respectively, see above). Moreover, this value is within the confidence interval for shuffled data (0.271 - 0.281) and is not significantly different from chance ($p \sim= 0.29$). We found similar effects for the other brain areas. Specifically, in vmPFC, the difference in $A_{idx}$ between the both medium-stakes and high-stakes and risky-safe is positive (0.051 and 0.018 for medium and high; $p<0.01$ for both; for this and the following, we used the bootstrap test described above). In pgACC, these numbers were 0.046 and 0.017 ($p<0.01$ for both); in PCC, the numbers were 0.041 and 0.009 ($p<0.01$ for both). In VS, the numbers were 0.038 and 0.007 ($p<0.01$ for medium and $p<0.05$ for high).

**Orthogonal risky and safe response subspaces can be transformed to be aligned**

Above, we demonstrated that the population subspaces for risky and safe offers are semi-orthogonal (see cartoon in **Figure 7A**). A primary question, then, is to what extent can these subspaces be transformed into a common subspace that would aid in their comparison? Such a common subspace could provide a mechanism for comparing offers of equal value (Yoo and Hayden, 2020). We reasoned that if the response subspaces could be transformed such that their hyperplanes were collinear, then the axis along which they become aligned is most likely the axis that describes their relative values (**Figure 7B**).

To investigate whether our data here obey these principles, we performed a canonical correlation analysis on the subspace loadings (i.e., the principal component projection weights, see **Methods;** Gallego et al., 2018; Susillo et al., 2015) for the first three principal components from the safe and risky offer response matrices. We also randomly shuffled the subspace loadings 1000 times and performed the same canonical correlation analysis. In OFC, we found that safe and equally valued medium-stakes offer response subspaces could be significantly aligned to correlation of r = 0.53 ($p$ = 0.023, bootstrap rank test). Safe and matched-high stakes risky offer response subspaces reached a maximal correlation of r = 0.61 ($p < 0.01$). We found similar results across the remaining five brain areas in our dataset ($p < 0.05$ in all cases, **Figure 8**). Specifically, we found that in all five brain areas, subspaces are linked. As above, we repeated these canonical correlation analyses, but for the two risky offers (medium and high stakes). They are also transformable. Specifically, in OFC, we find a correlation of r=0.62 (p = 0.011, bootstrap rank test). In vmPFC, we find a correlation of r=0.69 (p = 0.008, bootstrap rank test); in pgACC, we find a correlation of r=0.71 (p = 0.007, bootstrap rank test). In PCC, we find

a correlation of r=0.64 (p = 0.013, bootstrap rank test); finally, in VS, we find a correlation of r=0.72 (p = 0.007, bootstrap rank test).

One limitation of this analysis is that it is based on neural responses occurring during the first offer period only. That is, our analysis assumes implicitly that neural responses to the first offer will be identical to those for the second offer, and that cross-trial subspaces will allow comparison. This assumption can be tested. We therefore performed an analysis using neural responses to the second offers. Specifically, we performed a canonical correlation analysis on safe (offer 1) and equally valued middle-risk (offer 2), on safe (offer 1) and equally valued high-stakes (offer 2), as well as the reverse. In all cases, we found that the principle of transformability was preserved. Specifically, in OFC, we find a correlation of r=0.65 (p = 0.009, bootstrap rank test). In vmPFC, we find a correlation of r=0.68 (p = 0.01, bootstrap rank test); in pgACC, we find a correlation of r=0.70 (p = 0.006, bootstrap rank test). In PCC, we find a correlation of r=0.62 (p = 0.016, bootstrap rank test); finally, in VS, we find a correlation of r=0.67 (p = 0.010, bootstrap rank test).

## DISCUSSION

We find that neural responses to risky and safe offers use distinct codes in five core reward areas. Our task uses a single safe value and a dense sampling of 99 different risky values (1-99%), meaning we can use behavior to precisely infer equivalent risky and safe values. Indeed, we find that risky and safe options are encoded in distinct (semi-orthogonal) ensemble subspaces. The risky and safe subspaces, while different, are systematically related so that there is a ready transformation between them; this transformability means that a downstream area can straightforwardly compare the values of the two offers despite the differences in the codes used to represent them. We conjecture that this distinct but transformable code can allow the brain to achieve two conflicting goals, that is, to maintain a separate representation of different offer types, but also to allow for their direct comparison on a single scale.

Typically, neuroeconomic models are concerned with developing theories to explain how we compare disparate option types (Plassmann et al., 2007; Padoa-Schioppa, 2011; Levy and Glimcher, 2012). While abstraction is important, the more general problem faced by the brain is a bit more complex - it must use a system that simultaneously maintains options' features (so that they can be used to influence behavior if needed) while *also* allowing for comparison of dissimilar features. For example, circumstances may change rapidly so that surety is more or less valuable; the decision-maker must be able to selectively change the relative value of risky options. The use of linked but semi-orthogonal subspaces for qualitatively different feature types allows this goal to be accomplished quickly. Our results, then, support the notion that these ostensible value regions do contribute to evaluation and comparison processes in choice, but do so in a way that maintains information about the qualities of the options, in addition to their values.

Our results have some bearing on longstanding discussions about the nature of abstract value representation and their role in facilitating comparison (Padoa-Schioppa, 2011; Montague and Berns; Hayden and Niv, 2021; Platt and Glimcher, 1999; Padoa-Schioppa and Assad, 2006; Kennerley et al., 2009; Peters and Buchel, 2010; Kolling et al., 2016; So and Stuphorn, 2010; Raghuraman and Padoa-Schioppa, 2014; McCoy and Platt, 2005). Traditional neuroeconomics proceeds from the 'neuron doctrine,' which, in neuroeconomics, implies that comparison across offer types requires the existence of neurons whose responses are the same for equally valued offers of different types. By contrast, the 'population doctrine' allows for other forms of equivalence, including translatable but distinct subspaces (Elsayed et al., 2017; for reviews, see Saxena and Cunningham, 2019; Ebitz and Hayden, 2021). Equally valued risky and safe offers are a well-known example of offer types with qualitatively different features that can be readily compared, and for this reason, have long been the focus of studies about value representation (Levy et al., 2010; Tobler and Weber, 2014). Thus, our results both raise the possibility of and add empirical support for the idea that qualitatively different values can be compared in a way that occurs at the population level and does not require value neurons. To speculate a bit, subspace orthogonalization may offer more encoding flexibility than single neuron encodings because they can be rapidly adjusted to context without requiring rewiring or retraining; they may also be more robust to certain forms of error or degradation.

The concept of neural subspaces originates in motor cortex and can be used to explain how motor preparation is kept separate from action, so as to allow for fast responding while preventing precocious movement (Kaufman et al., 2014; Elsayed et al., 2016). We have argued that the same principles apply to core value regions, and that subspace orthogonalization allows for the separation of evaluation and comparison processes (Yoo and Hayden, 2020). The present

results suggest a new use for subspace partitioning - keeping track of two qualitatively different types of rewards in a format so that the details of their properties (here, risky vs. safe) are maintained, while also allowing for ready comparison in a downstream structure.

Orthogonalized subspaces are an emergent property of neurons with mixed selectivity. Mixed selectivity, or the simultaneous tuning to multiple features, is an important property of neurons in prefrontal and associated regions (Barak et al., 2013; Rigotti et al., 2013; Fusi et al., 2016). Mixed selectivity is the foundation of many useful properties, including feature binding, learning, abstraction, and generalization, and code morphing (Bernardi et al., 2020; Parthasarathy et al., 2017). Notably, a good deal of the analysis of physiological data assumes, either explicitly or tacitly, that neurons have singular tuning, not mixed selectivity. For example, approaches that divide neurons into specific categorical types tend to classify those neurons according to which feature drives them most strongly. By doing so, these approaches zero out those neurons' tuning for other parameters, and make the population appear more categorical than it really is. That work, however, risky misinterpretation of neural data, precisely because of the remarkable power of mixed selectivity (Fusi et al., 2016).

One highlight of our study is that we were able to directly compare responses in five brain regions in the same task. This fact means that we can ask questions about the unique functional properties of each region, such as whether different regions contribute differentially to evaluation, comparison, and action selection elements of choice. Here, we replicate our past findings showing that functional properties are largely qualitatively the same (Strait et al., 2015 and 2016; Maisson et al., 2021; Fine and Hayden, 2022). This does not necessarily prove these areas are strictly overlapping in their functions, nor does it prove that economic functions are purely distributed. Indeed, there is a quite a good deal of evidence of functional specialization in

660 the brain (Wilson et al., 2010; Rushworth et al., 2011; Passingham and Wise, 2012). Instead, we
661 propose that, to some extent, economic functions are a general feature of prefrontal cortex (and
662 related structures like PCC and VS), and that, for core economic functions, these regions
663 function as a hierarchy, rather than as a series of modules (Hunt and Hayden, 2017; Yoo and
664 Hayden, 2018 and 2021; Fine and Hayden, 2021; Maisson et al., 2021). Moreover, our results
665 here extend these previous ideas and suggest that these regions not only have similar coding
666 repertoires but use similar computational algorithms (here, judicious control of subspace) to
667 implement them.

668 Warren Weaver (1982) and Lola Lopes (1987) have both argued, in their work on
669 decision-making and risk, that it is a mistake to assume that two options with the same expected
670 value (including subjective expected value) are or should be treated the same way by decision-
671 makers. There are many natural situations in which a risky prospect is quite different from a safe
672 one, even if they are matched for subjective value. To give some examples, risky options elicit
673 emotions, and anxieties, promote learning, reward curiosity, generate error signals, require
674 maintenance of an eligibility trace, may be associated with exploratory rather than exploitative
675 states, require additional brain computations, and may in some cases differentially elicit
676 executive control (e.g. Loewenstein et al., 2001; McCoy and Platt, 2005; Slovic et al., 2007; Platt
677 and Huettel, 2008; Baraseghyan et al., 2013; So and Stuphorn, 2016). For these reasons, risky
678 and safe options ought to elicit at least somewhat distinct brain responses, even when they are
679 matched for subjective value. Our results show that in five brain areas, they elicit strikingly
680 different response patterns. This does not mean that an abstract single-neuron value code does
681 not exist in the brain; it may be found, for example, in a hierarchically later area, such as dorsal
682 anterior cingulate cortex (dACC, Cai and Padoa-Schioppa, 2012), or supplementary motor areas

(So and Stuphorn, 2010). Another possibility, however, is that the brain makes do without an abstract value subspace, and instead achieves comparison through alternative means (Vlaev et al., 2012; Yoo and Hayden, 2018; Hayden and Niv, 2021; Walasek and Brown, 2021).

686 **REFERENCES**


687 1. Azab, H., & Hayden, B. Y. (2017). Correlates of decisional dynamics in the dorsal
688 anterior cingulate cortex. PLoS biology, 15(11), e2003091.
689 2. Azab, H., & Hayden, B. Y. (2018). Correlates of economic decisions in the dorsal and
690 subgenual anterior cingulate cortices. European Journal of Neuroscience, 47(8), 979-993.
691 3. Azab, H., & Hayden, B. Y. (2020). Partial integration of the components of value in
692 anterior cingulate cortex. Behavioral Neuroscience, 134(4), 296.
693 4. Barak, O., Rigotti, M., & Fusi, S. (2013). The sparseness of mixed selectivity neurons
694 controls the generalization–discrimination trade-off. Journal of Neuroscience, 33(9),
695 3844-3856.
696 5. Binde, P. (2013). Why people gamble: A model with five motivational dimensions.
697 International Gambling Studies, 13(1), 81-97.
698 6. Blanchard, T. C., Piantadosi, S. T., & Hayden, B. Y. (2018). Robust mixture modeling
699 reveals category-free selectivity in reward region neuronal ensembles. Journal of
700 neurophysiology, 119(4), 1305-1318.
701 7. Blanchard, T. C., Wolfe, L. S., Vlaev, I., Winston, J. S., & Hayden, B. Y. (2014). Biases
702 in preferences for sequences of outcomes in monkeys. Cognition, 130(3), 289-299.
703 8. Cai, X., & Padoa-Schioppa, C. (2012). Neuronal encoding of subjective value in dorsal
704 and ventral anterior cingulate cortex. Journal of Neuroscience, 32(11), 3791-3808
705 9. Cisek, P., & Kalaska, J. F. (2010). Neural mechanisms for interacting with a world full of
706 action choices. Annual Review of Neuroscience, 33, 269-298.
707 10. Ebitz, R. B., & Hayden, B. Y. (2021). The population doctrine in cognitive neuroscience.
708 Neuron.

11. Elsayed, G. F., Lara, A. H., Kaufman, M. T., Churchland, M. M., & Cunningham, J. P. (2016). Reorganization between preparatory and movement population responses in motor cortex. Nature communications, 7(1), 1-15.

12. Farashahi, S., Donahue, C. H., Hayden, B. Y., Lee, D., & Soltani, A. (2019). Flexible combination of reward information across primates. Nature human behaviour, 3(11), 1215-1224.

13. Farashahi, S., Azab, H., Hayden, B., & Soltani, A. (2018). On the flexibility of basic risk attitudes in monkeys. Journal of Neuroscience, 38(18), 4383-4398.

14. Fine, J. M., & Hayden, B. Y. (2022). The whole prefrontal cortex is premotor cortex. Philosophical Transactions of the Royal Society B, 377(1844), 20200524.

15. Fusi, S., Miller, E. K., & Rigotti, M. (2016). Why neurons mix: high dimensionality for higher cognition. Current opinion in neurobiology, 37, 66-74.

16. Gallego, J. A., Perich, M. G., Miller, L. E., & Solla, S. A. (2017). Neural manifolds for the control of movement. Neuron, 94(5), 978-984.

17. Gallego, J. A., Perich, M. G., Naufel, S. N., Ethier, C., Solla, S. A., & Miller, L. E. (2018). Cortical population activity within a preserved neural manifold underlies multiple motor behaviors. Nature communications, 9(1), 1-13.

18. Glimcher, P. W., & Fehr, E. (Eds.). (2013). *Neuroeconomics: Decision making and the brain*. Academic Press.

19. Grupe, D. W., & Nitschke, J. B. (2013). Uncertainty and anticipation in anxiety: an integrated neurobiological and psychological perspective. Nature Reviews Neuroscience, 14(7), 488-501.

20. Hayden, B., Heilbronner, S., & Platt, M. (2010). Ambiguity aversion in rhesus macaques.

Frontiers in neuroscience, 4, 166.

21. Hayden, B. Y., & Niv, Y. (2021). The case against economic values in the orbitofrontal cortex (or anywhere else in the brain). Behavioral Neuroscience, 135(2), 192.
22. Heilbronner, S. R. (2017). Modeling risky decision-making in nonhuman animals: shared core features. Current opinion in behavioral sciences, 16, 23-29.
23. Heilbronner, S., & Hayden, B. (2013). Contextual factors explain risk-seeking preferences in rhesus monkeys. Frontiers in neuroscience, 7, 7.
24. Heilbronner, S., Hayden, B. Y., & Platt, M. (2011). Decision salience signals in posterior cingulate cortex. *Frontiers in Neuroscience*, 5(55).
25. Holt, C. A., & Laury, S. K. (2002). Risk aversion and incentive effects. American economic review, 92(5), 1644-1655.
26. Hunt, L. T., & Hayden, B. Y. (2017). A distributed, hierarchical and recurrent framework for reward-based choice. Nature Reviews Neuroscience, 18(3), 172-182.
27. Johnston, W. J., Palmer, S. E., & Freedman, D. J. (2020). Nonlinear mixed selectivity supports reliable neural computation. PLoS computational biology, 16(2), e1007544.
28. Kacelnik, A., & Bateson, M. (1996). Risky theories—the effects of variance on foraging decisions. American Zoologist, 36(4), 402-434.
29. Kahneman, T. & Tversky, A. (1979). Prospect Theory: An analysis of decisions under risk. *Econometric,*47(2).
30. Kaufman, M. T., Churchland, M. M., Ryu, S. I., & Shenoy, K. V. (2014). Cortical activity in the null space: permitting preparation without movement. Nature neuroscience, 17(3), 440-448.
31. Kennerley, S. W., Dahmubed, A. F., Lara, A. H., & Wallis, J. D. (2009). Neurons in the

755 frontal lobe encode the value of multiple decision variables. Journal of cognitive
756 neuroscience, 21(6), 1162-1178.
757 32. Kolling, N., Wittmann, M. K., Behrens, T. E., Boorman, E. D., Mars, R. B., &
758 Rushworth, M. F. (2016). Value, search, persistence and model updating in anterior
759 cingulate cortex. *Nature Neuroscience*, 19(10), 1280-1285.
760 33. Lau, B., & Glimcher, P. W. (2008). Value representations in the primate striatum during
761 matching behavior. *Neuron*, 58(3), 451-463.
762 34. Lerner, J. S., Li, Y., Valdesolo, P., & Kassam, K. S. (2015). Emotion and decision
763 making. Annual review of psychology, 66, 799-823.
764 35. Levy, D. J., & Glimcher, P. W. (2012). The root of all value: a neural common currency
765 for choice. Current opinion in neurobiology, 22(6), 1027-1038.
766 36. Levy, I., Snell, J., Nelson, A. J., Rustichini, A., & Glimcher, P. W. (2010). Neural
767 representation of subjective value under risk and ambiguity. *Journal of*
768 *neurophysiology*, *103*(2), 1036-1047.
769 37. Libby, A., & Buschman, T. J. (2021). Rotational dynamics reduce interference between
770 sensory and memory representations. Nature neuroscience, 24(5), 715-726.
771 38. Loewenstein, G. F., Weber, E. U., Hsee, C. K., & Welch, N. (2001). Risk as feelings.
772 Psychological bulletin, 127(2), 267.
773 39. Lopes, L. L. Between hope and fear: The psychology of risk. Adv. Exp. Soc. Psychol. 20,
774 255–295 (1987).
775 40. Maisson, D. J. N., Cash-Padgett, T. V., Wang, M. Z., Hayden, B. Y., Heilbronner, S. R.,
776 & Zimmermann, J. (2021). Choice-relevant information transformation along a
777 ventrodorsal axis in the medial prefrontal cortex. Nature communications, 12(1), 1-14.

778 41. McCoy, A. N., & Platt, M. L. (2005). Risk-sensitive neurons in macaque posterior
779 cingulate cortex. Nature neuroscience, 8(9), 1220-1227.
780 42. Padoa-Schioppa, C., & Assad, J. A. (2006). Neurons in the orbitofrontal cortex encode
781 economic value. Nature, 441(7090), 223-226.
782 43. Padoa-Schioppa, C. (2011). Neurobiology of economic choice: a good-based model.
783 Annual review of neuroscience, 34, 333-359.
784 44. Parthasarathy, A., Herikstad, R., Bong, J. H., Medina, F. S., Libedinsky, C., & Yen, S. C.
785 (2017). Mixed selectivity morphs population codes in prefrontal cortex. Nature
786 neuroscience, 20(12), 1770-1779.
787 45. Passingham, R. E., & Wise, S. P. (2012). The neurobiology of the prefrontal cortex:
788 anatomy, evolution, and the origin of insight (No. 50). Oxford University Press.
789 46. Paulus, M. P., & Angela, J. Y. (2012). Emotion and decision-making: affect-driven belief
790 systems in anxiety and depression. Trends in cognitive sciences, 16(9), 476-483.
791 47. Paxinos, G., Huang, X. F., & Toga, A. W. (2000). The rhesus monkey brain in stereotaxic
792 coordinates.
793 48. Plassmann, H., O'doherty, J., & Rangel, A. (2007). Orbitofrontal cortex encodes
794 willingness to pay in everyday economic transactions. Journal of neuroscience, 27(37),
795 9984-9988.
796 49. Platt, M. L., & Huettel, S. A. (2008). Risky business: the neuroeconomics of decision
797 making under uncertainty. Nature neuroscience, 11(4), 398-403.
798 50. Platt, M. L., & Glimcher, P. W. (1999). Neural correlates of decision variables in parietal
799 cortex. Nature, 400(6741), 233-238.
800 51. Peters, J., & Büchel, C. (2010). Neural representations of subjective reward

801 value. *Behavioural Brain Research*, 213(2), 135-141.
802 52. Raghuraman, A. P., & Padoa-Schioppa, C. (2014). Integration of multiple determinants in
803 the neuronal computation of economic values. Journal of Neuroscience, 34(35), 11583-
804 11603.
805 53. Raposo, D., Kaufman, M. T., & Churchland, A. K. (2014). A category-free neural
806 population supports evolving demands during decision-making. Nature neuroscience,
807 17(12), 1784-1792.
808 54. Reyna, V. F., & Huettel, S. A. (2014). Reward, representation, and impulsivity: A
809 theoretical framework for the neuroscience of risky decision making.
810 55. Rigotti, M., Barak, O., Warden, M. R., Wang, X. J., Daw, N. D., Miller, E. K., & Fusi, S.
811 (2013). The importance of mixed selectivity in complex cognitive tasks. Nature,
812 497(7451), 585-590.
813 56. Rushworth, M. F., Noonan, M. P., Boorman, E. D., Walton, M. E., & Behrens, T. E.
814 (2011). Frontal cortex and reward-guided learning and decision-making. Neuron, 70(6),
815 1054-1069.
816 57. Saxena, S., & Cunningham, J. P. (2019). Towards the neural population doctrine. *Current*
817 *opinion in neurobiology*, *55*, 103-111.
818 58. Slovic, P., Finucane, M. L., Peters, E., & MacGregor, D. G. (2007). The affect heuristic.
819 European journal of operational research, 177(3), 1333-1352.
820 59. Slovic, P., Finucane, M. L., Peters, E., & MacGregor, D. G. (2004). Risk as analysis and
821 risk as feelings: Some thoughts about affect, reason, risk, and rationality. Risk Analysis:
822 An International Journal, 24(2), 311-322.
823 60. So, N. Y., & Stuphorn, V. (2010). Supplementary eye field encodes option and action

824 value for saccades with variable reward. Journal of Neurophysiology, 104(5), 2634-2653.
825 61. Strait, C. E., Blanchard, T. C., & Hayden, B. Y. (2014). Reward value comparison via
826 mutual inhibition in ventromedial prefrontal cortex. Neuron, 82(6), 1357-1366.
827 62. Strait, C. E., Sleezer, B. J., & Hayden, B. Y. (2015). Signatures of value comparison in
828 ventral striatum neurons. PLoS biology, 13(6), e1002173.
829 63. Strait, C. E., Sleezer, B. J., Blanchard, T. C., Azab, H., Castagno, M. D., & Hayden, B.
830 Y. (2016). Neuronal selectivity for spatial positions of offers and choices in five reward
831 regions. Journal of neurophysiology, 115(3), 1098-1111.
832 64. Sussillo, D., Churchland, M. M., Kaufman, M. T., & Shenoy, K. V. (2015). A neural
833 network that finds a naturalistic solution for the production of muscle activity. Nature
834 neuroscience, 18(7), 1025-1033.
835 65. Tang, C., Herikstad, R., Parthasarathy, A., Libedinsky, C., & Yen, S. C. (2020).
836 Minimally dependent activity subspaces for working memory and motor preparation in
837 the lateral prefrontal cortex. Elife, 9, e58154.
838 66. Tobler, P. N., & Weber, E. U. (2014). Valuation for risky and uncertain choices.
839 In *Neuroeconomics* (pp. 149-172). Academic Press.
840 67. Yoo, S. B. M., Hayden, B. Y., & Pearson, J. M. (2021). Continuous decisions.
841 Philosophical Transactions of the Royal Society B, 376(1819), 20190664.
842 68. Yoo, S. B. M., & Hayden, B. Y. (2020). The transition from evaluation to selection
843 involves neural subspace reorganization in core reward regions. Neuron, 105(4), 712-724.
844 69. Yoo, S. B. M., & Hayden, B. Y. (2018). Economic choice as an untangling of options
845 into actions. Neuron, 99(3), 434-447.
846 70. Walasek, L., & Brown, G. (2021). Incomparability and Incommensurability in Choice:

No Common Currency of Value?. psyarXiv.

71. Wang, M. Z., & Hayden, B. Y. (2021). Latent learning, cognitive maps, and curiosity. Current Opinion in Behavioral Sciences, 38, 1-7.

72. Wang, M. Z., Hayden, B. Y., & Heilbronner, S. R. (2022). A structural and functional subdivision in central orbitofrontal cortex. Nature communications, 13(1), 3623.

73. Weaver, W. Lady luck: The theory of probability. (Dover Press, 1982).

74. Wilson, C. R., Gaffan, D., Browning, P. G., & Baxter, M. G. (2010). Functional localization within the prefrontal cortex: missing the forest for the trees?. Trends in neurosciences, 33(12), 533-540.

75. Vlaev, I., Chater, N., Stewart, N., & Brown, G. D. (2011). Does the brain calculate value?. Trends in cognitive sciences, 15(11), 546-554.

**Author contributions:**

Conceptualization: DM, JF, SY, JZ, BH

Methodology: DM, BH, JF

Investigation: TC, MW, BS, JF

Visualization: DM

Supervision: JZ, BH, JF

Writing – original draft: DM, JZ, BH

Writing – review & editing: DM, JF, SY, JZ, BH

**Data and materials availability:**

All analyses were performed using stock Matlab functions and no custom code was generated. Data used for all reported analyses are available on Dryad (https://doi.org/10.5061/dryad.ttdz08kx9).

# Figure Captions

**Figure 1.** *Task, targeted structures, behavior, and calculation of equivalent risky and safe values.* **(A)** Structure of our Risky Choice Task (Strait et al., 2014). Task consists of 400 ms offer 1 presentation, a 600 ms blank period, a 400 ms offer 2 presentation, another 600 ms blank period, then a fixation spot and, on fixation, reappearance of both offers and choice (indicated by saccade). For each offer, the magnitude of the associated reward (stakes) is indicated by the bottom color (green, high or blue, medium) of the stimulus. The probability of being rewarded is indicated by its size. **(B)** Anatomical positions of our brain regions of interest: OFC (purple), vmPFC (purple-pink), pgACC (pink), PCC (gold), and VS (grey). **(C)** Likelihood of choosing the first offer as a function of its value relative to the second (specifically, for signed value difference). Sigmoid fits of raw binary data shown (see **Methods**). Gray lines: individual subjects; black line: group average. In this and subsequent panels, a horizontal dashed line indicates the indifference point (the point at which choices are 50/50). **(D)** Likelihood of choosing a safe option as a function of the probability of the risky option for medium (blue) and high (green) stakes offers. All data were analyzed on a subject-by-subject basis, so only data for one example subject (subject B) are shown. Other subjects showed similar patterns. Vertical black lines indicate the probability used as the SV-equivalence point for the subject (the arrow points to the indifference point for medium- (blue) and high-stakes (green) risky offers). **(E)** Same as D, except data are separated for left and right offers. Side of presentation does not affect choice much. **(F)** Same as D, except data are separated by trials that were in the first (early) or second (late) half of a session.

**Figure 2.** Indifference points for each of our five subjects individually. Y-axis shows the probability (from 0% to 100%) associated with a risky offer that is equivalent in subjective value, as assessed by preference indifference. Data are separated by whether the offer is medium (left) or high (right) stakes). Subjects generally prefer risky offers, as indicated by the fact that points tend to be on the lower y-values of the graph. Preferences are lower for high stakes than for medium stakes, indicating that subjects took account of stakes. Subject indifferent to stakes would show no difference between the medium and high stakes conditions. Error bars are not shown because, due to the large number of trials, they are smaller than the bars.

**Figure 3.** *Responses of single neurons.* This figure shows the average responses of sample neurons to safe and risky offers of differing values, as well as the average response similarity rates. **(A)** Peristimulus time histogram from mean firing rates of sample neuron vmPFC.70. Each line indicates the average response across offers of a given risk profile (grey: all safe offers; blue: all medium-stakes risky offers; green: all large stakes risky offers). The grey shaded box indicates the 1-second period from which the 500-ms epoch 1 analysis window was extracted, where the onset of the first offer is time-locked to zero seconds. **(B)** This is a plot of data collected from a sample neuron pgACC.17, which showed a response to safe offers that was statistically different from the response to equivalent risky offers. Depicted are the average responses to medium-stakes (left; blue) and high-stakes (right; green), separated by probability ranges of 0.05. The black point indicates the average response of the given neuron to safe offers (error bars denote the SEM across responses to safe offers). The diagonal black line indicates a fitted regression line, showing positive monotonic tuning. **(C-E)** Same as (B) but demonstrating

sample cell responses to an assortment of medium- and high-stakes offers from across all target areas.

**Figure 4.** *Overlapping populations encode both safe and risky offers.* **(A)** Conceptual schematic indicating the analytical approach for identifying functional subpopulations. The right panels propose the slope of a best fit line describing the encoding formats for each population type. **(B)** For OFC (area 13), left panel: scatter plots represent the absolute degree of encoding of safe offers plotted against that of the medium-stakes risky offers for each neuron (each point corresponds to data from one neuron). The black line indicates the line of best fit for the correlation between the absolute degree of encoding of safe and risky offers across neurons. Red bars indicate the 99% confidence interval for the line of best fit. The inlaid numbers indicate the Pearson correlation coefficient (r) and the significance (p). Right panel: same as left, but comparing safe offer encoding and high-stakes risky offer encoding. **(C-F)** Same as B, but for vmPFC (area 14), pgACC (area 32), PCC (area 29/31), and ventral striatum, respectively.

**Figure 5.** *Safe and equivalently valued risky offers are readily decoded.* Decodability for safe and risky offers is high in all areas for both medium- and high-stakes gambles. Shuffled data refers to decodability of randomly assigned safe/risky labels to neural responses that are completely shuffled across trials and cells. Bars represent the mean decoding accuracy across out of sample cross-validations Error bars indicate the standard error across cross-validations. The horizontal grey line indicates the chance decoding accuracy level.

**Figure 6.** *Subspace Alignment between Safe and Equivalent Risky Offers.* **(A)** For each structure,

we performed a principal components analysis and projected the resulting dimensions for medium (top row) and high (bottom row) stakes offers onto the corresponding ones for safe offers. If subspaces for the risky and safe codes were the same (or collinear), then the projections would match. Instead, the projections for the risky offers (lighter lines) are consistently lower than the control values (the projections for the safe offers, darker lines). In other words, the fact that the lighter (risky) lines are lower than the darker (safe) ones indicates that the two types of offers are encoded in semi-orthogonal subspaces. **(B)** This finding is summarized in panel B, which gives a summary of the difference between the average shuffle alignment index and either the safe-medium (left bar) or the safe-high (right bar) alignment index. Error bars indicate the standard error across computed differences. Note that these data represent the difference between the measured value and the shuffle alignment index; this means that data with higher shuffle alignment indices will appear lower as a result.

**Figure 7.** *Subspace rotational transformations and canonical correlations.* **(A)** A cartoon demonstrating an example of orthogonal (left) and aligned (right) hyperplanes projected onto the first three dimensions (cf. Yoo and Hayden, 2020). **(B)** A procedural schematic demonstrating the process of linear algebraic rotational transformation, to align subspaces by maximizing the Pearson's correlation coefficient via canonical correlation.

963 **Figure 8.** Scatter plots, for each structure, of projections of medium and safe offer responses (top
964 row) and high and safe offer responses (bottom row) projected into safe offer response
965 subspaces. Plotted are projections onto the first 2 principal component loadings for the original
966 (darker circles) and transformed (lighter pluses) projections. The inlaid coefficients are the
967 Pearson's correlation coefficients between projections on the first 2 principal component loading
968 for the original (r1) and transformed (r2) responses. The lines denote the best-fit line for the
969 corresponding correlation coefficients (r1: darker line; r2: lighter line).
970