**Attractor dynamics reflect decision confidence in macaque prefrontal cortex**

Siyu Wang1\*, Rossella Falcone1,2\*, Barry Richmond1, Bruno B. Averbeck1†

*1Laboratory of Neuropsychology, National Institute of Mental Health, National Institutes of Health, Bethesda MD, 20892-4415*

*2Current affiliation: Leo M. Davidoff Department of Neurological Surgery, Albert Einstein College of Medicine Montefiore Medical Center, Bronx,NY, 10461.*

*\*equal contribution*

†Correspondence to:

Bruno B. Averbeck, Ph.D.

Laboratory of Neuropsychology, NIMH/NIH

Building 49 Room 1B80

49 Convent Drive MSC 4415

Bethesda, MD 20892-4415

E-mail: bruno.averbeck@nih.gov

Abstract

Decisions are made with different degrees of consistency, and this consistency can be linked to the confidence that the best choice has been made. Theoretical work suggests that attractor dynamics in networks can account for choice consistency, but how this is implemented in the brain remains unclear. Here, we provide evidence that the energy landscape around attractor basins in population neural activity in prefrontal cortex reflects choice consistency. We trained two rhesus monkeys to make accept/reject decisions based on pretrained visual cues that signaled reward offers with different magnitudes and delays-to-reward. Monkeys made consistent decisions for very good and very bad offers, but decisions were less consistent for intermediate offers. Analysis of neural data showed that the attractor basins around patterns of activity reflecting decisions had steeper landscapes for offers that led to consistent decisions. Therefore, we provide neural evidence that energy landscapes predict decision consistency, which reflects decision confidence.

Keywords: population dynamics; decision certainty; prefrontal cortex; neurophysiology

**INTRODUCTION**

Decisions are made with varying levels of certainty. For example, we may be certain about which political candidate we prefer, but less certain about which entree to order at a restaurant. When we are certain about a choice, we are more likely to make the same choice consistently. However, when we are uncertain, our choices are less consistent, and we may be more likely to change our mind before making a final decision 1. There is a substantial body of literature looking at theoretical, behavioral, and neural aspects of decision certainty, also referred to as decision confidence 2-4. It has been shown that, across several species, prefrontal and parietal cortical areas represent decision certainty 5-11. While these studies have established that multiple brain areas contribute to decision certainty, they have not examined the network computational mechanisms that drive variability in decision making under different levels of certainty.

A prominent family of models that account for flexible decisions are the two-state attractor models. In these models, decisions are made when network activity settles into one of two attractor basins 12-19. The models operationalize the intuitive idea that decision related neural activity can be modeled as an object frictionally sliding on a landscape characterized by two valleys separated by a hill 20. The valleys represent the attractor basins. The ball starts on top of the hill and is pushed into one of the valleys by evidence in favor of one of the choices. The decision corresponds to the valley in which the ball ends up 21. These models account for a wide range of behavior including perceptual and value based decision making 15-17, and working memory 21. Change-of-mind and variability in decision making is accounted for in these models by assuming that transient stimuli or noise pushes neural activity from one attractor basin (i.e., valley) to another 17,22-24. Theoretical work using artificial network simulations suggests that choice consistency can be accounted for by the shape of the energy landscape around the attractor basins. When energy landscapes have steep sides and deep basins, corresponding to a high hill between the valleys, decisions will be made consistently, because it is hard to drive activity from one basin to another. However, when energy landscapes are relatively flat and attractor basins are shallow, decisions will be made less consistently, because it is easier to drive activity between attractor basins. It has been shown, for example, that reinforcement learning can increase the energy separation between the attractor basins, as optimal choices are learned, which can lead to increased consistency of choices 25,26. When attractor basins are shallower, before values have been learned, choices are less consistent 25.

Despite these advances in theory, less work has examined directly the structure of the dynamics in neural activity that underlies variability in decision making. Dynamical systems models have been used to understand neural activity in motor control 27-30, decision making 31-34, representation of head direction in rodent anterior thalamic nuclei 35 and the encoding of aggression in the hypothalamus 36. Most studies have focused on the average trajectories that underlie movements and choices as they unfold. Recently, however, researchers have examined flow fields and energy landscapes underlying these mean trajectories 35 and explicitly examined attractor dynamics by fitting dynamical system models to neural data 34,36. Other work has examined changes in single-cell dynamics, by comparing time constants across brain areas 37-39. Studies on the computations that underlie decision making have focused on evidence integration or drift-diffusion models 40-42. These models have led to substantial insights, but they do not incorporate fixed-point dynamics, which may be an important component of neural decision mechanisms. To the best of our knowledge, there has been no direct neural evidence that the geometry of attractor basins, which reflect the energy landscape that drives neural activity, predicts decision certainty. In this work, we investigated population dynamics of prefrontal neurons using high channel-count Utah-array recordings (768 total channels; 43) while monkeys performed a decision-making task. We trained two monkeys to choose between rejecting and accepting offers of different reward sizes and delay. By linking neural attractor dynamics to decision certainty, we provide evidence that the energy landscape in prefrontal cortex predicts decision certainty in behavior.

**RESULTS**

We trained rhesus monkeys to perform a decision-making task 44(Fig. 1A). In this task, animals had to reject or accept an offer by releasing a bar at different times. In each trial, animals were presented with one of nine possible reward offers, signaled by a visual stimulus. Each offer was a combination of reward size (2, 4 or 6 drops) and delay (1, 5 or 10 seconds; Fig. 1B). Monkeys began a trial by touching a bar and holding it for 500ms. Then a red square was presented at the center of the screen for 500ms. After that, one of the nine visual cues (Fig. 1B) was presented at the center of the screen behind the central red square. The central square stayed red for a random period of between 1500 and 3500ms before it turned purple. To reject the offer, the monkey had to release the bar before the square turned purple. After an offer was rejected, a new trial started, with a new cue randomly selected from the set of possible cues. To accept the offer, the monkey had to hold the bar until the red square turned purple and then release it between 200ms and 1200ms after the presentation of the purple square. After an offer was accepted, the purple square turned green, and the designated number of drops was delivered after the delay indicated by the visual cue. The visual cue was turned off during reward delivery. The two monkeys completed on average 848 (Monkey V) and 761 (Monkey W) valid trials per session. Data from 16 sessions (8 for each monkey) were analyzed.

We computed the proportion of times that monkeys accepted each offer, (Fig. 1C). Monkeys almost always accepted offers with large rewards and short delays (e.g., 6 drops and 1s delay), and almost always rejected offers with small rewards and long delays (e.g., 2 drops and 10s delay). For other offers which had intermediate values (e.g., 4 drops and a 5s delay), the acceptance probability was intermediate. The decision consistency for each offer was quantified by the entropy of choices (see Methods). Entropy is a natural measure of uncertainty in probability distributions 45. Lower entropy characterized consistent choices for both the best (large reward, short delay) and worst (small reward, long delay) offers. Entropy was higher for the least consistent choices for intermediate offers (Fig. 1D).

We also analyzed whether decision consistency was reflected in reaction time (RT). Reaction times for reject decisions (from cue onset to bar release) peaked between 350 and 450ms for the two monkeys (Fig. 1E, Supplementary Fig. S1A). We found that the average reaction time correlated significantly with decision consistency for each offer in reject decisions (p < 0.001 for both monkeys, Fig. 1F). Reaction times were slower when decision consistency was smaller. This correlation may not entirely reflect decision consistency, but rather motivation. As a negative control, we analyzed response time for accept decisions (from purple dot to bar release). Accept response time also significantly correlated with decision consistency (p = 0.001 for monkey V, p < 0.001 for monkey W, Supplementary Fig. S1B). Since the decision, presumably, has already been made when the dot turns purple (Supplementary Fig. S1C), this correlation should not reflect the decision-making calculation. Rather it may reflect the speed of simple visual motor reaction time. To control for the effect of motivation, we performed an additional combined regression analysis (see Methods) which included both p(accept), the probability of accepting an offer which serves as a quantification of motivation, and choice entropy which serves as a measure of decision consistency. Our results showed that both p(accept) and choice entropy significantly predict reaction time for monkey W (p < 0.01 for p(accept), p < 0.05 for entropy), and only choice entropy statistically predict reaction time for monkey V (p > 0.05 for p(accept), p < 0.05 for entropy). This further suggests that decision consistency is reflected in reaction times, after controlling for motivation.

**Neurons in PFC show activity associated with choice, reward size, and delay time.**

While the two monkeys performed the behavioral task, we recorded the extracellular activity of neurons from the prefrontal cortex using a bilateral implant of 8 multielectrode Utah arrays (10x10 layout, 96 electrodes each), 4 in each hemisphere (Fig. 1G). We recorded on average 326.9 (Monkey V) and 527.5 (Monkey W) neurons simultaneously per session for the two monkeys. The recorded neurons were evenly distributed across left and right hemispheres.

We aligned the neural activity to the onset of the visual offer cues in each trial. Then we counted the number of spikes within a sliding window (50ms width, 50ms steps) from 1000ms before cue onset to 2000ms after cue onset for each trial and neuron. We performed a sliding window multivariate ANOVA analysis on spike counts (Fig. 2A, Supplementary Fig. S2). The results showed that a significant proportion of neurons responded to both the features of the offer (delay time, reward size, and their interaction) and the choice to either accept or reject the offer.

Next, we carried out principal component analysis on the simultaneously recorded neurons from each session. The PCA was done on the mean spiking activity for each offer cue and each choice type (accept vs reject) in 50ms bins from 0 to 1000ms after cue onset. The first 20 principal components accounted for about 65% of the variance (Supplementary Fig. S3A). We found that we could decode trial-by-trial choice (accept vs. reject), using a linear SVM with 10-fold cross validation, at close to 90% accuracy based on the neural population activity projected into the 20-dimensional PCA subspace (Fig. 2B). We were also able to decode information about reward size and delay from the 20-D neural subspace (Supplementary Fig. S3B, C). Adding more dimensions did not increase the decoding performance (Supplementary Fig. S4). The decoding of the two monkeys’ choices were above baseline from 200ms after cue onset.

**Neural trajectories and energy landscapes in the choice dimension**

We examined the neural dynamics underlying decisions in a one-dimensional neural subspace using activity in the simultaneously recorded neurons from each session from each animal. This allowed us to visualize the way in which population activity, including trial-to-trial variability in population activity, tended to evolve as decisions developed. We began by reconstructing a numerical approximation to the energy landscape in this 1-D space (see Methods). To carry out this analysis we first projected the 20-dimensional population activity in each trial and each time bin, , onto the one-dimensional choice dimension defined by the SVM (Fig. 3A, B). This gave us a point in the 1-D space, , at each time t. We then computed time derivatives, , in the 1-D space (Fig. 3B). This was done for all times considered, and all trials. We then binned the derivatives in the 1-D space, by finding all points that fell into each spatial bin along the 1-D axis (Fig. 3C) and calculated the average derivative for all points starting at that fell into each bin. This provided an estimate, on average, of how the neural activity evolved in this 1-D space, as a function of location in the 1-D space. These derivatives are estimates of the vector field for the dynamics underlying population activity (46, see Methods). The vector field characterizes the forces, driven by cortical networks and their inputs, acting on activity as it evolves through time.

In the next step we spatially integrated these average derivatives, across the 1-D space, to generate an estimate of the potential function that drove neural activity (Fig. 3D, illustration, Fig. 3E, data). On average, neural activity evolves from high potential energy to lower potential energy, similar to an object frictionally sliding on a landscape. Points with small spatial derivatives led to flat spots in the energy landscape, because when the neural activity fell into these locations it tended to stay close to these locations, on average. Correspondingly, points with large spatial derivatives led to steep spots on the energy landscape, because neural activity that fell into these locations tended to evolve away from them quickly, in the direction of decreasing energy (Fig. 3D). Stated another way, population activity tends to move from a high potential position to a low potential position, following the gradients.

We examined the energy landscapes computed separately for the accepted and rejected trials. For these analyses we averaged across cue conditions. We merged the landscapes at the SVM decision point (position 0), such that the landscape on one side of the decision point was derived from analyses of the accept trials, and the landscape on the other side was derived from analyses of the reject trials. By plotting these energy landscapes over time, our results showed that the attractor basins for accept and reject choices were deepest at around the time of the decision (~300 to 600ms after cue onset) (Fig. 3F). We overlayed the average activity in this space, for accept and reject decisions, and it could be seen that the average approximately followed the low point on the landscape across time (Fig. 3F).

While an energy landscape defined as the integral of the flow vectors can be computed in the 1-D subspace, true energy functions for higher dimensional neural subspaces may not exist. The existence of an energy function implies that the integration of gradients in the vector field between two points along different paths lead to the same quantity, or correspondingly that the integral along a closed path is zero. For 3 dimensions, the existence of an energy function requires the curl of the field (see Methods) to be zero at all points in space. Although we do not have access to the curl at all points in space, we empirically estimated the curl around the mean population trajectory based on our local estimate of the recurrent dynamics in a 3-dimensional neural subspace (see Methods). The estimated curl was not significantly different from zero at most time points (Supplementary Fig. S5). This does not prove the existence of an energy function. But we did not observe statistically significant non-zero curls. A non-zero curl would violate the assumption of the existence of an energy function.

**Decision consistency was correlated with the steepness of the attractor basins**

We further investigated how the energy landscape in the 1-D subspace changed as a function of decision consistency. We tested the hypothesis that shallower attractor basins at the time of decision were associated with lower decision consistency. During decision formation, neural activity will tend to frictionally slide downhill on these landscapes, towards the low points (attractor basins or valleys) that represent the decisions. However, neural activity is also subject to noise, and the noise can drive the neural activity out of one valley and into another. Our hypothesis, therefore, suggests that deeper attractor basins will lead to more consistent decisions because variability in neural activity is less likely to stochastically drive population activity out of the deeper basins.

We computed the energy landscape separately for each of the nine cues, averaged across trials in which both accept and reject choices were made. These energy landscapes were computed for each session and then averaged across sessions. We found that the landscapes had steeper attractor basins following presentation of cues that had high decision consistency (accept decisions for large reward short delay, or rejection decisions for small reward long delay) and shallower attractor basins following cues that had low decision consistency (intermediate reward and delay) (Fig. 4).

Most of the landscapes appeared to have only one basin as opposed to two basins. This is partially because we have more data for the basin that corresponds to the more favorable choice by the animal for each cue. For example, the population activity rarely visits the negative portion of the 1-D subspace for cues which animals choose to accept all the time, thus we do not have an estimate of the reject basin. Furthermore, because our observation of population position is a noisy readout of the true position, this noise could potentially flatten the observed landscape. Partially because of this, there appears to be 1 basin even for cues that had low decision consistency in which both positive and negative parts of the subspace are visited. However, when we reconstructed the basins separately for these cues, for accept and reject choices, it could be seen that they had separate basins for each choice (Supplementary Fig. S6). We further examined this by comparing the fit of one and two basin models, and found that our data was better fitted by a 2-basin model compared to a 1-basin model (1-basin model: AIC = 2533.8; 2-basin model: AIC = 2515.0). The minima of the reconstructed basins also shifted outward for consistent choices (Fig. 4), which was also reflected in the mean choice trajectories (Supplementary Fig. S7). These reconstructed landscapes are, therefore, consistent with our hypothesis. However, the differences across conditions could reflect differences in noise for the different conditions as well as the evidence in favor of each choice, as opposed to differences in attractor dynamics. Therefore, in the next analysis, we fit models to estimate the neural dynamics, as well as the effect of evidence, derived from a probabilistic model, and the overall noise level.

We fit dynamical systems models to compare three hypotheses that can account for the neural activity during decision formation. The simplest model assumes that activity in prefrontal cortex reflects only the decision, without reflecting either the evidence used to arrive at the decision, or neural dynamics. The second hypothesis is that the activity reflects a (strict) drift diffusion process. Drift-diffusion processes assume perfect integration of evidence, and no underlying fixed-point dynamics. Finally, we examined the hypothesis that choices are driven by both evidence accumulation and fixed-point dynamics. These three hypotheses can be compared by fitting a single model and examining the parameters. Specifically, we fit a model which predicted the time dependent neural activity as a function of a cue dependent evidence term, , derived from the behavior (Supplementary Fig. S8A, also see Methods), and a cue and decision dependent dynamics term, :

In this model is time, refers to one of the nine different cues and indexes choice. The variables , , and are twelve free parameters, fit by nonlinear regression, to predict the time-dependent neural activity, , and is zero-mean Gaussian white noise. The variable characterizes the fixed-point dynamics. We refer to as the retraction coefficient. A larger reflects a steeper attractor basin. The variable characterizes the strength of the evidence in favor of each choice, and the variable characterizes the fixed-point of the undriven system for each choice (i.e., reject vs. accept). The undriven fixed point is an estimate of the mean position in state space to which the activity would go when the evidence term, , was 0 in the model. The evidence terms , were constants defined by fitting the choice consistency and reaction time behavioral data (Supplementary Fig. S8A). The variable was estimated using three different versions of drift diffusion model-fitting methods. The results were similar between the three model-fits (Supplementary Fig. S9, S10). Note that if prefrontal activity represents a drift-diffusion process, the coefficients would be 0, and if the activity reflects only the choice without dynamics, the coefficients would be 1 and the coefficient would be 0.

When we examined the coefficients of the model, we found that the choice process, , was strongly represented around the reject decision reaction time and reflected the decision (Fig. 5A). Further, we found that the activity was affected by the evidence in favor of each choice, (Fig. 5B). The evidence term was non-significant during the baseline hold, and peaked early, while the decision was being formed, but before the mean reaction time (Fig. 5B). Finally, we found that there were also significant fixed-point dynamics that did not reflect a perfect integration process, but that did reflect decision consistency (Fig. 5C, Supplementary Fig. S10). We further examined whether the depth of the attractor, characterized by the coefficients on the recurrent dynamics, , reflected choice entropy (Fig. 5D, E). We found that, at the time of choice, the dynamics were correlated with the choice entropy across cues (Fig. 5E). This correlation also was non-significant during the baseline hold, and became significant during choice formation, after the cues were shown (Fig. 5D). When the monkeys made consistent, low entropy decisions, either accept or reject, the retraction coefficients were larger (Fig. 5C, Supplementary Fig. S10). Thus, we found evidence for fixed-point dynamics in choice-related prefrontal cortex activity. We further found that the depth of the attractor basins reflected choice confidence, consistent with our numerical reconstruction of the energy landscape (Fig. 4) because a larger retraction coefficient reflects a steeper energy function. Thus, consistent decisions had deeper attractor basins, and inconsistent decisions had shallower basins.

We further examined the relationship between evidence accumulation and neural activity by fitting a variation of the model (Supplementary Fig. S11), in which we estimated a single constant, , for each cue, without explicitly including the evidence term, :

.

Thus, in this model we fit an evidence constant directly to the neural activity. We then compared the cue dependent constant, , fit to the neural data, to the cue dependent evidence term derived from the behavior, , and found that they correlated significantly, following onset of the cues (Supplementary Fig. S8B). This further supports the hypothesis that prefrontal neural activity reflects both evidence accumulation and fixed-point dynamics.

To exclude the possibility that the number of accept or reject choices for a particular cue drives the differences in estimates of retraction coefficients based on our model fitting procedure, we also conducted a control analysis. We refit the model after matching the number of trials between the consistent and inconsistent choice conditions. Our analysis suggests that the retraction coefficient was significantly larger even after matching the number of trials in the different conditions (Supplementary Fig. S12). We also controlled for the effect of ongoing change-of-mind on our model estimates. To consider changes of mind during the choice period, we predicted the current choice preference in each time bin (instead of using the final choice of the animal) and attributed the activity classified in this way to accept or reject decisions for the analysis. By repeating our linear dynamical system analysis using moment-by-moment estimates of choice, instead of the final choice, we accounted for the potential differences in the amount of “changes-of-mind” between different cue conditions. In this alternative analysis, our results still hold (Supplementary Fig. S13).

To further support our findings, we examined a 1-D choice dimension determined by an alternative classifier (see Methods, and Supplementary Fig. S14A-E). The estimated retraction coefficients, using choice dimensions defined by the SVM classifier vs the alternative classifier, were significantly correlated (Supplementary Fig. S14F). We also repeated our analysis in a 3-dimensional subspace spanned by the 1-D choice dimension and the first two principal components of the residual activity orthogonal to the choice dimension. Eigenvalues of the recurrent matrix were computed, in place of the 1-D retraction coefficients. Consistent with our 1-D analysis, we found similar correlations between choice entropy and the eigenvalues in the 3-D analysis. Larger eigenvalues, which reflect a steeper basin, were associated with low choice entropy and high choice consistency (Supplementary Fig. S15A-C). Moreover, the vector connecting the fitted undriven fixed points for accept and reject decisions in the 3-D analysis aligns with the 1-D choice dimension, and the cosine similarity between the two peaked at the time of choice (Supplementary Fig. 15D). Thus, analyses in 3-dimensions lead to results that agree with our findings in 1-dimension.

We also examined how individual neurons’ responses to value affect the observed attractor dynamics. A separate sliding window multivariate ANOVA analysis with two features, choice and value, was carried out to examine individual neuron’s tuning to offer value. Behaviorally, offer value is reflected in the probability that the monkeys accept each offer. The results showed that the percentages of neurons that respond positively to value (accept probability) were 46.5% for Monkey V, and 37.8% for monkey W. Neurons that respond more to better offers also have higher loading weights on the 1-D choice dimension (Supplementary Fig. S16C). As a sanity check, we examined mean firing rates of individual neurons, and showed that negative preferring neurons have enhanced firing rates for reject offers, whereas positive preferring neurons have enhanced firing rates for accept offers (Supplementary Fig. S16A-B). Next, we repeated our dynamical system analysis using only positive, only negative or a mixture of positive and negative value neurons. For each session, the number of neurons for the three groups was selected to be the minimum between the number of positive and negative neurons to control for differences in population size. A random sample was taken for the other group and the mixture group. We observed that attractor retraction coefficients were larger for consistent offers for all three groups (Supplementary Fig. S17). This suggests that attractor dynamics do not differentially reflect activities of positive or negative offer preferring neurons.

**Reaction time was correlated by the steepness of the attractor basins**

Since the derivatives in steeper basins are larger, we hypothesized that reaction times would also be smaller for steeper basins, under the assumption that choices are triggered when neural activity reaches a defined location in state space, which we refer to as the choice zone 14,29,47. In theory, it takes fewer steps for neural activity to reach the choice zone in steeper basins with higher retraction coefficients. Similar to the analysis of choice consistency vs retraction coefficients, we correlated the retraction coefficients with reaction time across cues. We found that the retraction coefficients significantly correlated with reaction time at the time of decision for reject choices in both monkeys (Fig. 6A-C).

As a negative control, we also computed the correlation between retraction coefficients extracted at the average reaction time for reject choices and reaction times for accept choices. Our results showed only a weak correlation between retraction coefficients and reaction time for accept choices in Monkey W (p = 0.02), and none in Monkey V (p > 0.05) (Supplementary Fig. S18). This supports the suggestion that the correlation between reaction time for reject decisions and retraction coefficients is not merely due to the behavioral correlation between reaction time and decision consistency because the correlation is weaker for accept decisions. However, this only holds statistically in monkey V. Together, this supports our hypothesis that steeper attracter basins (higher ) are associated with shorter reaction times.

**DISCUSSION**

In this paper, we investigated the neural underpinnings of choice consistency by analyzing dynamics in prefrontal population activity during decision making. We found that attractor basins were shallower following cues that signaled intermediate value offers, which also led to lower decision consistency. Correspondingly, attractor basins were deeper following cues that signaled high or low value offers, that led to higher decision consistency. Further, by fitting linear dynamical systems models to the neural data we found that the retraction coefficients, which characterize the energy landscape around the attractor, correlated significantly with both the entropy of choice consistency and reaction time at the time of decision. Larger retraction coefficients, which are consistent with steeper energy landscapes, were associated with cues that elicited consistent decisions. Our results provide neural evidence that attractor dynamics predict decision consistency.

To visualize the dynamics of neural activity, we empirically estimated the energy landscape in the state space of population activity. However, the existence of a true energy function is non-trivial and requires curls of the vector field to be zero (or in higher dimensions, the partial derivatives of the dynamics equation have to be symmetric) at all spatial points in the neural state space. We estimated the curl around the mean population trajectory based on our fitted linear dynamical systems model in a 3-dimensional subspace. We did not observe curls statistically different from zero. However, this does not prove the existence of an energy function. First, failure to reject a null hypothesis does not prove the null hypothesis, thus failure to detect a non-zero curl does not prove the curl to be zero. Second, we only estimated the curl around the mean population trajectory and do not have global estimates of the curl at other spatial locations of the state space. Whether different neural systems, and different systems under different conditions, have energy functions is an interesting question. Considering whether networks in the brain have potential functions opens up several interesting neural and behavioral questions. Given plasticity rules, like spike-timing dependent plasticity, one wonders whether it is possible to ever return to the same brain state. Future studies are needed to directly examine the curl in different neural systems under different tasks.

We fit locally linear dynamical systems models, that let us compare three hypotheses, that can account for the decision related activity in prefrontal cortex. The simplest hypothesis was that prefrontal activity would represent the choice, independent of the evidence that supported that choice, or dynamics related to the choice-formation process. The second hypothesis was that prefrontal activity would reflect a drift-diffusion process, that integrated evidence in favor of each choice. The third hypothesis was that prefrontal activity would reflect both evidence accumulation and fixed-point dynamics, consistent with network models that instantiate decision processes. We found that our dynamical systems model results supported the third hypothesis. Prefrontal activity reflected both the evidence in favor of each choice, as well as fixed-point dynamics. Furthermore, the dynamics were also related to choice consistency, such that choices made with greater consistency were characterized by deeper attractor basins. The dynamical systems models also agreed with direct numerical reconstructions of the energy landscapes, which also reflected the choice consistency. This association between energy landscapes and choice consistency has been predicted with computational models 23,25,26. Intuitively it follows from the assumption that steeper landscapes will more likely lead to consistent decisions given the inherent stochasticity in neural activity. Variability in neural activity is less likely to drive neural population activity out of a deeper attractor basin. This association between choice consistency and the energy landscape has not been shown previously in neural activity, to our knowledge.

In our case, monkeys made accept and reject decisions based on the utility of an offer. Because reject decisions led to a random draw from the possible offers in the next trial, each trial’s offer could be compared to the average offer and the delay to the next trial. We found, consistent with previous work on this paradigm 44,48, that some offers were consistently rejected because they were well below average, and some offers were consistently accepted because they were well above average. Offers with intermediate values, however, were sometimes accepted and sometimes rejected. Because internal representations of offer values are stochastic, and choice probability is a function of value differences 8,49, this leads to situations where choices are made with varying levels of consistency. Choice consistency can be related to decision certainty and decision confidence, which is the subjective probability that the chosen option has the highest value 3,5,10,11,50-52. Consistently accepted or rejected offers reflect high confidence choices, because the probability that the monkey is making the highest value decision is closer to 1, whereas inconsistently accepted or rejected offers reflect low confidence choices, because the probability that the monkey is making the highest value decision is closer to 0.5.

Several mechanisms have been proposed for representation of decision confidence. For example, drift-diffusion models (DDMs) have been developed to account for information integration in perceptual inference, value-based decision making, and memory recall tasks, and used to model decision confidence 8,50,53-55. In perceptual inference variants of these models, it is assumed that an internal process is integrating noisy information from an external stimulus 50. In value-based decision-making variants of these models it is assumed that an internal process is integrating a noisy internal representation of offer value 8,56. Across these models, the integration process can be related to a Bayesian posterior probability that a given choice is correct, and these posteriors can be used to define decision confidence. These models, and the associated Bayesian posteriors, provide good descriptions of the formal mechanism that gives rise to confidence, in many 57, but not all 58,59 conditions. However, they do not specify the neural mechanisms by which the brain implements these formal computations.

Network attractor models for perceptual and value-based decision making have also been developed 13-16,18,22,24. These models reproduce changes in mean population activity with decision confidence 16,18. They do this by integrating activity from pools of neurons representing chosen and unchosen options. Because neural activity levels are higher for options when they are chosen with high confidence, confidence can be read out by integrating across the chosen and unchosen option pools. This suggests a mechanism by which activity levels related to certainty about chosen options 18 can be converted directly to choice-independent confidence estimates 11. Changes in the energy landscape of most of these networks, when choices are driven by different levels of information, have not been explored. Recent work, however, has examined changes in neural dynamics when reinforcement learning drives decisions from uncertain to more certain 25. This study found that before learning which of two option was more valuable, in an RL task 60, choice options were not well separated across the energy landscape. However, after learning, the options were better separated.

In empirical work, previous studies have shown that certainty, or probability distributions over sensory variables, can be estimated by decoding probabilistic population codes 57,61-64, or through related sampling models 65. When choice is decoded from neural activity, posterior probabilities of choices have also been shown to reflect decision consistency, and therefore average decision confidence 3,5,66,67. Single neurons in OFC have been shown to directly encode decision confidence, independent of chosen actions, in odor and auditory discrimination tasks 10,11. Thus, mean activity levels have been shown to reflect decision confidence.

We also found that the fixed-point of the dynamics, related to the mean choice-related activity, associated with options that were chosen more consistently was represented farther from the decision boundary. The fixed-point of the dynamics is given by . Therefore, if we know the fixed point of the undriven () dynamics, , decision confidence can be read out by observing the fixed-point into which the system settles, at the decision time. (Note that also scales with and therefore has to scale sub-linearly for this to hold.) This distinguishes the dynamical representation of evidence, from a DDM with a bound, because activity level (the position of the diffusion particle) and time have to be available to estimate the Bayesian posterior in a DDM. Furthermore, in our model activity settles into a fixed point that represents the evidence in a time-independent way. In a DDM, activity continues to increase over time, unless one assumes a sticky bound, and therefore it never directly represents the evidence. Whether perceptual decision-making processes would also show dynamics similar to what we have seen in the context of value-based decision making is an open question.

Because the fixed point reflects the evidence, confidence would not have to be estimated from the energy landscape, i.e., one would not have to compute an estimate of the energy landscape on a single trial to estimate confidence, even though the dynamics are correlated with decision confidence. The dynamics, however, reflect the energy required to change a decision. Flatter dynamics reflect the fact that choices are less consistent when they are associated with cues that predict intermediate value offers. The dynamics we have observed, therefore, likely reflect the changeability of a deliberative choice process when cues represent intermediate value offers, or a commitment to a decision when cues represent very good or very bad offers. Due to the correlational nature of our analysis, future perturbation studies are required to examine the causal role of attractor dynamics in change-of-mind and decision confidence. Related dynamics may also underlie observed variability in population neural choice representations seen in other decision-making tasks 68,69. Because dynamics are also correlated with mean distances in decision space, this also raises the question of whether dynamics and posterior probabilities are linked, in an obligatory way, in biological and artificial networks.

**Conclusion**

Significant progress, both theoretical and experimental, has been made in our understanding of how populations of neurons code information about sensory variables, decisions, and behavior 70,71. There is also a substantial body of theoretical work that defines how computations in networks can underlie complex behavior 15,20,35,72-77. However, there has been less experimental progress in this area, except in motor control 78,79 and hippocampal circuitry 35,77. In the present study, we have begun to make progress on linking computations in networks to population representations during decision making. We found that neural dynamics reflected choice consistency and reaction times, when monkeys accepted and rejected offers with different values. High and low value offers were accepted or rejected consistently, with short reaction times. The neural dynamics in these situations were characterized by deep attractor basins and correspondingly large retraction coefficients in a dynamical systems model. Intermediate value offers, on the other hand, were characterized by inconsistent choices, and longer reaction times. The neural dynamics for these offers had shallower attractor basins and correspondingly smaller retraction coefficients. Thus, the dynamic landscape of population neural activity in prefrontal cortex reflects choice commitment during value-based decision making. These results extend significantly the insights into behavior that can be obtained by studying populations of simultaneously recorded neurons, and directly link theoretical work on network computations to experimental results.

**Acknowledgements**

This work was supported by the intramural research program of NIMH (ZIA MH002928 BA). The authors thank Cary Robinson and Yi Wei for their assistance in spike sorting.

**Author Contributions**

R.F. conducted the experiment and collected data. S.W. analyzed the data. S.W. and B.A. wrote the manuscript, with input from R.F. and B.R.

**Competing Interests**

The authors declare no competing interests.

**Figure Legends/Captions**

**Figure 1** – Behavioral task, behavior, and recording sites. A. Temporal-discounting task 44. In each trial, the animals first touched a bar, after which a red dot was presented. Next, a cue, which indicated the reward size (drops) and delay, was shown. If the animals released the bar before the red dot turned purple, the trial was considered rejected and, in the next trial, a new randomly selected cue was presented. If they held the bar until the red dot turned purple, and then released the bar between 200 and 1200 ms, the trial was considered accepted, and the indicated drops of liquid reward were delivered after the indicated delay. B. The nine visual cues used to indicate reward size and reward delay. C. Probability of accepting the given offer as a function of each offer of reward size and delay for the two monkeys. N = 8 for each animal. Data are presented as mean values +/- SEM. D. Entropy of p(accept) for each offer. N = 8 for each animal. Data are presented as mean values +/- SEM. Note entropy is high when offers are not consistently accepted or rejected, and low when offers are consistently accepted or rejected. E. Reaction time distributions for the two monkeys for reject trials only. F. Reaction time vs. decision entropy for reject trials. Note that individual points correspond to individual sessions. For monkey V, r = 0.82, p < 0.001; for monkey W, r = 0.82, p < 0.001. G. Recording locations in left and right hemispheres in each animal 80. PS: Principal Sulcus; Arc: Arcuate Sulcus; CS: Central Sulcus.

**Figure 2** – Neuron populations in PFC encode choice, and the offer value. Error bar indicates +/- 1 s.e.m. n = 16 for both animals or n = 8 for each animal. A. Fraction of single neurons across both monkeys and recording sessions that significantly encoded each indicated task factor assessed with a multivariate ANOVA (p < 0.05). B. Decoding accuracy for accept vs. reject decisions from the 20-D PCA space defined for each session, as a function of time relative to cue onset.

**Figure 3**. Reconstruction of energy landscape. A. The choice dimension defined by the SVM classifier, i.e., the dimension that is orthogonal to the separating hyperplane in the high-dimensional neural space (20-D PCA space). B. (cartoon) Activity, , in 50 ms time bins, in high-dimensional neural space is projected onto the 1-D choice dimension. This results in a time series of projected activity, . Activity was analyzed in 50 ms bins, so ms. Panel B illustrates the projected temporal trajectory of a single trial. C. (cartoon) Illustration of the temporal derivatives for 3 trials and the computation of mean derivatives in each spatial bin in the choice dimension. D. (cartoon) Integration over the mean derivatives in each spatial bin gives us the energy landscape or potential function. At each spatial location, the derivative of the energy landscape equals the mean derivative derived in C. E. The computed energy landscape for reject decisions at an example time point. F. The evolution of the energy landscapes, within the 1-D choice dimension, over time for both monkeys for accept and reject decisions. Position 0 is the SVM decision boundary and time 0 is cue onset. The light green line indicates the average activity for reject trials, and the red line indicates the average activity for accept trials. For the portion of the state space that was unvisited, there is no gradient or potential estimate, and it is left blank. The choice-dimension for all bins was the same and defined at the mean reaction time in each session.

**Figure 4** – Numerical estimate of energy landscapes for each cue. Data are presented as mean values +/- SEM. A-B. The landscapes have qualitatively shallower attractor basins following presentation of cues that had low decision consistency for the two monkeys. The lines are separated by the probability of accepting the offer, and correspondingly, positive values in the choice dimensions reflect accept choices. We also examined landscapes separated by accept and reject decisions for the uncertain decisions (Supplementary Fig. S6).

**Figure 5** – Neural activity reflects evidence accumulation and fixed-point dynamics. Error bars represent standard error of the mean. For plots that show data separately for each monkey, N = 8 as SEM was computed across sessions for each animal. A. Choice related activity, characterized as the undriven fixed points, over time. Note the darker color curve is accept decision and lighter color curve is reject decision, for each monkey. B. Strength of evidence-driven neural activity over time, coefficient *b*. The curve peaks at 375 ms for monkey V and 325 ms for monkey W. C. Retraction coefficients are higher for more certain decisions for both reject and accept choices. Certain accept = 90%+ vs. uncertain 50-90%. And certain reject 0-10% vs. uncertain 10-50% accept choices. Further, coefficients differ from both 0 and 1 and therefore there are fixed-point dynamics, and the system is not a perfect integrator. D. Correlation between retraction coefficient for each cue, and the behavioral entropy of that cue. A high retraction coefficient parameter corresponds to a steeper basin. E. The retraction coefficients correlated significantly with decision consistency at the time of decision (i.e., mean reaction time in each session for each monkey). For monkey V, r = -0.56, p < 0.001; for monkey W, r = -0.64, p < 0.001.   
\* = 0.05, \*\* = 0.01, \*\*\* = 0.001

**Figure 6** – Retraction coefficients vs reaction time. For plots that show data separately for each monkey, N = 8 as SEM was computed across sessions for each animal. A. The correlations between retraction coefficients and reaction time for reject decisions over time. B. The correlations between retraction coefficients at median time of reject decision time and reaction time for reject decisions (monkey V, r = -0.82, p < 0.001). C. Same as B for monkey W (monkey W, r = -0.59, p < 0.001).**References**

1 Resulaj, A., Kiani, R., Wolpert, D. M. & Shadlen, M. N. Changes of mind in decision-making. *Nature* **461**, 263-266 (2009). <https://doi.org:10.1038/nature08275>

2 Fleming, S. M. & Frith, C. D. *The cognitive neuroscience of metacognition*. (Springer, 2014).

3 Kepecs, A. & Mainen, Z. F. A computational framework for the study of confidence in humans and animals. *Philos Trans R Soc Lond B Biol Sci* **367**, 1322-1337 (2012). <https://doi.org:10.1098/rstb.2012.0037>

4 Pouget, A., Drugowitsch, J. & Kepecs, A. Confidence and certainty: distinct probabilistic quantities for different goals. *Nat Neurosci* **19**, 366-374 (2016). <https://doi.org:10.1038/nn.4240>

5 Kiani, R. & Shadlen, M. N. Representation of confidence associated with a decision by neurons in the parietal cortex. *Science* **324**, 759-764 (2009). <https://doi.org:10.1126/science.1169405>

6 Fleming, S. M., Huijgen, J. & Dolan, R. J. Prefrontal contributions to metacognition in perceptual decision making. *J Neurosci* **32**, 6117-6125 (2012). <https://doi.org:10.1523/JNEUROSCI.6489-11.2012>

7 Fleming, S. M., Weil, R. S., Nagy, Z., Dolan, R. J. & Rees, G. Relating introspective accuracy to individual differences in brain structure. *Science* **329**, 1541-1543 (2010). <https://doi.org:10.1126/science.1191883>

8 De Martino, B., Fleming, S. M., Garrett, N. & Dolan, R. J. Confidence in value-based choice. *Nat Neurosci* **16**, 105-110 (2013). <https://doi.org:10.1038/nn.3279>

9 Lebreton, M., Abitbol, R., Daunizeau, J. & Pessiglione, M. Automatic integration of confidence in the brain valuation signal. *Nat Neurosci* **18**, 1159-1167 (2015). <https://doi.org:10.1038/nn.4064>

10 Masset, P., Ott, T., Lak, A., Hirokawa, J. & Kepecs, A. Behavior- and Modality-General Representation of Confidence in Orbitofrontal Cortex. *Cell* **182**, 112-126 e118 (2020). <https://doi.org:10.1016/j.cell.2020.05.022>

11 Kepecs, A., Uchida, N., Zariwala, H. A. & Mainen, Z. F. Neural correlates, computation and behavioural impact of decision confidence. *Nature* **455**, 227-231 (2008). <https://doi.org:10.1038/nature07200>

12 Prat-Ortega, G., Wimmer, K., Roxin, A. & de la Rocha, J. Flexible categorization in perceptual decision making. *Nat Commun* **12**, 1283 (2021). <https://doi.org:10.1038/s41467-021-21501-z>

13 Machens, C. K., Romo, R. & Brody, C. D. Flexible control of mutual inhibition: a neural model of two-interval discrimination. *Science* **307**, 1121-1124 (2005).

14 Wong, K. F., Huk, A. C., Shadlen, M. N. & Wang, X. J. Neural circuit dynamics underlying accumulation of time-varying evidence during perceptual decision making. *Front Comput Neurosci* **1**, 6 (2007). <https://doi.org:10.3389/neuro.10.006.2007>

15 Wang, X. J. Probabilistic decision making by slow reverberation in cortical circuits. *Neuron* **36**, 955-968 (2002).

16 Rolls, E. T., Grabenhorst, F. & Deco, G. Decision-making, errors, and confidence in the brain. *J Neurophysiol* **104**, 2359-2374 (2010). <https://doi.org:10.1152/jn.00571.2010>

17 Deco, G., Rolls, E. T., Albantakis, L. & Romo, R. Brain mechanisms for perceptual and reward-related decision-making. *Prog Neurobiol* **103**, 194-213 (2013). <https://doi.org:10.1016/j.pneurobio.2012.01.010>

18 Insabato, A., Pannunzi, M., Rolls, E. T. & Deco, G. Confidence-related decision making. *J Neurophysiol* **104**, 539-547 (2010). <https://doi.org:10.1152/jn.01068.2009>

19 Wong, K. F. & Wang, X. J. A recurrent network mechanism of time integration in perceptual decisions. *J Neurosci* **26**, 1314-1328 (2006). <https://doi.org:10.1523/JNEUROSCI.3733-05.2006>

20 Amit, D. J. *Modeling brain function : the world of attractor neural networks*. (Cambridge University Press, 1989).

21 Inagaki, H. K., Fontolan, L., Romani, S. & Svoboda, K. Discrete attractor dynamics underlies persistent activity in the frontal cortex. *Nature* **566**, 212-217 (2019). <https://doi.org:10.1038/s41586-019-0919-7>

22 Atiya, N. A. A., Rano, I., Prasad, G. & Wong-Lin, K. A neural circuit model of decision uncertainty and change-of-mind. *Nat Commun* **10**, 2287 (2019). <https://doi.org:10.1038/s41467-019-10316-8>

23 Averbeck, B. B. Pruning recurrent neural networks replicates adolescent changes in working memory and reinforcement learning. *Proc Natl Acad Sci U S A* **119**, e2121331119 (2022). <https://doi.org:10.1073/pnas.2121331119>

24 Atiya, N. A. A., Huys, Q. J. M., Dolan, R. J. & Fleming, S. M. Explaining distortions in metacognition with an attractor network model of decision uncertainty. *PLoS Comput Biol* **17**, e1009201 (2021). <https://doi.org:10.1371/journal.pcbi.1009201>

25 Marton, C. D., Schultz, S. R. & Averbeck, B. B. Learning to select actions shapes recurrent dynamics in the corticostriatal system. *Neural Netw* **132**, 375-393 (2020). <https://doi.org:10.1016/j.neunet.2020.09.008>

26 Finkelstein, A. *et al.* Attractor dynamics gate cortical information flow during decision-making. *Nat Neurosci* **24**, 843-850 (2021). <https://doi.org:10.1038/s41593-021-00840-6>

27 Churchland, M. M. *et al.* Neural population dynamics during reaching. *Nature* **487**, 51-56 (2012). <https://doi.org:10.1038/nature11129>

28 Shenoy, K. V., Sahani, M. & Churchland, M. M. Cortical control of arm movements: a dynamical systems perspective. *Annu Rev Neurosci* **36**, 337-359 (2013). <https://doi.org:10.1146/annurev-neuro-062111-150509>

29 Shenoy, K. V., Kaufman, M. T., Sahani, M. & Churchland, M. M. A dynamical systems view of motor preparation: implications for neural prosthetic system design. *Prog Brain Res* **192**, 33-58 (2011). <https://doi.org:10.1016/B978-0-444-53355-5.00003-8>

30 Ebitz, R. B. & Hayden, B. Y. The population doctrine in cognitive neuroscience. *Neuron* **109**, 3055-3068 (2021). <https://doi.org:10.1016/j.neuron.2021.07.011>

31 Okazawa, G., Hatch, C. E., Mancoo, A., Machens, C. K. & Kiani, R. Representational geometry of perceptual decisions in the monkey parietal cortex. *Cell* **184**, 3748-3761 e3718 (2021). <https://doi.org:10.1016/j.cell.2021.05.022>

32 Mante, V., Sussillo, D., Shenoy, K. V. & Newsome, W. T. Context-dependent computation by recurrent dynamics in prefrontal cortex. *Nature* **503**, 78-84 (2013). <https://doi.org:10.1038/nature12742>

33 Zhao, Y., Yates, J. L., Levi, A. J., Huk, A. C. & Park, I. M. Stimulus-choice (mis)alignment in primate area MT. *PLoS Comput Biol* **16**, e1007614 (2020). <https://doi.org:10.1371/journal.pcbi.1007614>

34 Galgali, A. R., Sahani, M. & Mante, V. Residual dynamics resolves recurrent contributions to neural computation. *Nat Neurosci* **26**, 326-338 (2023). <https://doi.org:10.1038/s41593-022-01230-2>

35 Chaudhuri, R., Gercek, B., Pandey, B., Peyrache, A. & Fiete, I. The intrinsic attractor manifold and population dynamics of a canonical cognitive circuit across waking and sleep. *Nat Neurosci* **22**, 1512-1520 (2019). <https://doi.org:10.1038/s41593-019-0460-x>

36 Nair, A. *et al.* An approximate line attractor in the hypothalamus encodes an aggressive state. *Cell* **186**, 178-193 e115 (2023). <https://doi.org:10.1016/j.cell.2022.11.027>

37 Soltani, A., Murray, J. D., Seo, H. & Lee, D. Timescales of Cognition in the Brain. *Curr Opin Behav Sci* **41**, 30-37 (2021). <https://doi.org:10.1016/j.cobeha.2021.03.003>

38 Spitmaan, M., Seo, H., Lee, D. & Soltani, A. Multiple timescales of neural dynamics and integration of task-relevant signals across cortex. *Proc Natl Acad Sci U S A* **117**, 22522-22531 (2020). <https://doi.org:10.1073/pnas.2005993117>

39 Hart, E. & Huk, A. C. Recurrent circuit dynamics underlie persistent activity in the macaque frontoparietal network. *Elife* **9** (2020). <https://doi.org:10.7554/eLife.52460>

40 Gold, J. I. & Shadlen, M. N. Banburismus and the brain: decoding the relationship between sensory stimuli, decisions, and reward. *Neuron* **36**, 299-308 (2002).

41 Drugowitsch, J., Moreno-Bote, R., Churchland, A. K., Shadlen, M. N. & Pouget, A. The cost of accumulating evidence in perceptual decision making. *The Journal of neuroscience : the official journal of the Society for Neuroscience* **32**, 3612-3628 (2012). <https://doi.org:10.1523/JNEUROSCI.4010-11.2012>

42 Meister, M. L., Hennig, J. A. & Huk, A. C. Signal multiplexing and single-neuron computations in lateral intraparietal area during decision-making. *J Neurosci* **33**, 2254-2267 (2013). <https://doi.org:10.1523/JNEUROSCI.2984-12.2013>

43 Mitz, A. R. *et al.* High channel count single-unit recordings from nonhuman primate frontal cortex. *J Neurosci Methods* **289**, 39-47 (2017). <https://doi.org:10.1016/j.jneumeth.2017.07.001>

44 Falcone, R. *et al.* Temporal Coding of Reward Value in Monkey Ventral Striatal Tonically Active Neurons. *J Neurosci* **39**, 7539-7550 (2019). <https://doi.org:10.1523/JNEUROSCI.0869-19.2019>

45 Abramson, N. *Information Theory and Coding*. 1 edn, (McGraw-Hill, 1963).

46 Strogatz, S. H. *Nonlinear dynamics and Chaos : with applications to physics, biology, chemistry, and engineering*. (Addison-Wesley Pub., 1994).

47 Berlemont, K. & Nadal, J. P. Perceptual Decision-Making: Biases in Post-Error Reaction Times Explained by Attractor Network Dynamics. *J Neurosci* **39**, 833-853 (2019). <https://doi.org:10.1523/JNEUROSCI.1015-18.2018>

48 Minamimoto, T., La Camera, G. & Richmond, B. J. Measuring and modeling the interaction among reward size, delay to reward, and satiation level on motivation in monkeys. *J Neurophysiol* **101**, 437-447 (2009). <https://doi.org:10.1152/jn.90959.2008>

49 Lee, E., Seo, M., Dal Monte, O. & Averbeck, B. B. Injection of a dopamine type 2 receptor antagonist into the dorsal striatum disrupts choices driven by previous outcomes, but not perceptual inference. *J Neurosci* **35**, 6298-6306 (2015). <https://doi.org:10.1523/JNEUROSCI.4561-14.2015>

50 Drugowitsch, J., Mendonca, A. G., Mainen, Z. F. & Pouget, A. Learning optimal decisions with confidence. *Proc Natl Acad Sci U S A* **116**, 24872-24880 (2019). <https://doi.org:10.1073/pnas.1906787116>

51 Khalvati, K., Kiani, R. & Rao, R. P. N. Bayesian inference with incomplete knowledge explains perceptual confidence and its deviations from accuracy. *Nat Commun* **12**, 5704 (2021). <https://doi.org:10.1038/s41467-021-25419-4>

52 Lak, A. *et al.* Orbitofrontal cortex is required for optimal waiting based on decision confidence. *Neuron* **84**, 190-201 (2014). <https://doi.org:10.1016/j.neuron.2014.08.039>

53 Drugowitsch, J. Becoming Confident in the Statistical Nature of Human Confidence Judgments. *Neuron* **90**, 425-427 (2016). <https://doi.org:10.1016/j.neuron.2016.04.023>

54 Balsdon, T., Wyart, V. & Mamassian, P. Confidence controls perceptual evidence accumulation. *Nat Commun* **11**, 1753 (2020). <https://doi.org:10.1038/s41467-020-15561-w>

55 Charles, L. & Yeung, N. Dynamic sources of evidence supporting confidence judgments and error detection. *J Exp Psychol Hum Percept Perform* **45**, 39-52 (2019). <https://doi.org:10.1037/xhp0000583>

56 Lee, E., Seo, M., Dal Monte, O. & Averbeck, B. B. Injection of a Dopamine Type 2 Receptor Antagonist into the Dorsal Striatum Disrupts Choices Driven by Previous Outcomes, But Not Perceptual Inference. *Journal of Neuroscience* (2015).

57 Geurts, L. S., Cooke, J. R. H., van Bergen, R. S. & Jehee, J. F. M. Subjective confidence reflects representation of Bayesian probability in cortex. *Nat Hum Behav* **6**, 294-305 (2022). <https://doi.org:10.1038/s41562-021-01247-w>

58 Li, H. H. & Ma, W. J. Confidence reports in decision-making with multiple alternatives violate the Bayesian confidence hypothesis. *Nat Commun* **11**, 2004 (2020). <https://doi.org:10.1038/s41467-020-15581-6>

59 Adler, W. T. & Ma, W. J. Comparing Bayesian and non-Bayesian accounts of human confidence reports. *Plos Computational Biology* **14** (2018). <https://doi.org:ARTN> e1006572

10.1371/journal.pcbi.1006572

60 Seo, M., Lee, E. & Averbeck, B. B. Action selection and action value in frontal-striatal circuits. *Neuron* **74**, 947-960 (2012). <https://doi.org:10.1016/j.neuron.2012.03.037>

61 Ma, W. J., Beck, J. M., Latham, P. E. & Pouget, A. Bayesian inference with probabilistic population codes. *Nat Neurosci* **9**, 1432-1438 (2006).

62 Beck, J., Ma, W. J., Latham, P. E. & Pouget, A. Probabilistic population codes and the exponential family of distributions. *Prog Brain Res* **165**, 509-519 (2007). <https://doi.org:10.1016/S0079-6123(06)65032-2>

63 Beck, J. M. *et al.* Probabilistic population codes for Bayesian decision making. *Neuron* **60**, 1142-1152 (2008). <https://doi.org:10.1016/j.neuron.2008.09.021>

64 Zhang, K., Ginzburg, I., McNaughton, B. L. & Sejnowski, T. J. Interpreting neuronal population activity by reconstruction: unified framework with application to hippocampal place cells. *J Neurophysiol* **79**, 1017-1044 (1998). <https://doi.org:10.1152/jn.1998.79.2.1017>

65 Haefner, R. M., Berkes, P. & Fiser, J. Perceptual Decision-Making as Probabilistic Inference by Neural Sampling. *Neuron* **90**, 649-660 (2016). <https://doi.org:10.1016/j.neuron.2016.03.020>

66 Averbeck, B. B., Sohn, J. W. & Lee, D. Activity in prefrontal cortex during dynamic selection of action sequences. *Nat Neurosci* **9**, 276-282 (2006).

67 Bartolo, R., Saunders, R. C., Mitz, A. R. & Averbeck, B. B. Dimensionality, information and learning in prefrontal cortex. *PLoS Comput Biol* **16**, e1007514 (2020). <https://doi.org:10.1371/journal.pcbi.1007514>

68 Rich, E. L. & Wallis, J. D. Decoding subjective decisions from orbitofrontal cortex. *Nature Neuroscience* **19**, 973-980 (2016). <https://doi.org:10.1038/nn.4320>

69 Enel, P., Wallis, J. D. & Rich, E. L. Stable and dynamic representations of value in the prefrontal cortex. *Elife* **9** (2020). <https://doi.org:ARTN> e54313

10.7554/eLife.54313

70 Averbeck, B. B., Latham, P. E. & Pouget, A. Neural correlations, population coding and computation. *Nat Rev Neurosci* **7**, 358-366 (2006). <https://doi.org:10.1038/nrn1888>

71 Bartolo, R., Saunders, R. C., Mitz, A. R. & Averbeck, B. B. Information-Limiting Correlations in Large Neural Populations. *J Neurosci* **40**, 1668-1678 (2020). <https://doi.org:10.1523/JNEUROSCI.2072-19.2019>

72 Sussillo, D. & Abbott, L. F. Generating coherent patterns of activity from chaotic neural networks. *Neuron* **63**, 544-557 (2009). <https://doi.org:10.1016/j.neuron.2009.07.018>

73 Khona, M. & Fiete, I. R. Attractor and integrator networks in the brain. *Nat Rev Neurosci* **23**, 744-766 (2022). <https://doi.org:10.1038/s41583-022-00642-0>

74 Sejnowski, T. J. On the stochastic dynamics of neuronal interaction. *Biol Cybern* **22**, 203-211. (1976).

75 Zhang, K. Representation of spatial orientation by the intrinsic dynamics of the head-direction cell ensemble: a theory. *J Neurosci* **16**, 2112-2126 (1996). <https://doi.org:10.1523/JNEUROSCI.16-06-02112.1996>

76 Gardner, R. J. *et al.* Toroidal topology of population activity in grid cells. *Nature* **602**, 123-128 (2022). <https://doi.org:10.1038/s41586-021-04268-7>

77 Yoon, K. *et al.* Specific evidence of low-dimensional continuous attractor dynamics in grid cells. *Nat Neurosci* **16**, 1077-1084 (2013). <https://doi.org:10.1038/nn.3450>

78 Russo, A. A. *et al.* Motor Cortex Embeds Muscle-like Commands in an Untangled Population Response. *Neuron* **97**, 953-966 e958 (2018). <https://doi.org:10.1016/j.neuron.2018.01.004>

79 Russo, A. A. *et al.* Neural Trajectories in the Supplementary Motor Area and Motor Cortex Exhibit Distinct Geometries, Compatible with Different Classes of Computation. *Neuron* **107**, 745-758 e746 (2020). <https://doi.org:10.1016/j.neuron.2020.05.020>

80 Bartolo, R. & Averbeck, B. B. Prefrontal Cortex Predicts State Switches during Reversal Learning. *Neuron* **106**, 1044-1054 e1044 (2020). <https://doi.org:10.1016/j.neuron.2020.03.024>