Biological profits of irrational computations in the orbitofrontal cortex

Juliette Bénon[1] and Jean Daunizeau[1]

[1]Paris Brain Institute, France

Abstract

Making good decisions is essential for survival and success, yet humans and animals often exhibit perplexing irrational decision-making whose biological origin remains poorly understood. Recent theoretical work suggests that some forms of irrational decisions may arise from limited coding precision or metabolic budget in individual orbitofrontal neurons. Here, we consider the alternative possibility that systematic errors in decision-relevant computations are the inevitable consequence of the internal connectivity structure within orbitofrontal networks, which was molded under more distal biological constraints. We first trained cohorts of artificial neural networks to perform rational decision-relevant computations. Remarkably, they exhibited most electrophysiological coding properties of orbitofrontal neurons recorded in monkeys engaged in a preference-based decision task. We then distorted their internal connectivity to reproduce monkeys' irrational choices. This induced systematic interferences in decision-relevant computations that generalize across individuals, at both the behavioral and neural level. Importantly, irrational networks also display enhanced behavioral resilience to neural loss when compared to their rational counterparts. This suggests that irrational behavior may be the incidental outcome of distal evolutionary pressure on the tolerance to orbitofrontal circuit's damage.

1 Introduction

People and animals arguably act, in some circumstances, against their own interest. Why does irrational behavior persist, despite its potential costs to survival and fitness? Standard decision theory posits that rational decisions rely on estimating and comparing the expected value of each available alternative option in the choice set. Thus, irrational behavior may emerge from the covert mechanisms through which the brain constructs, maintains or compares option values. Decades of work in human and non-human primates show that these computational processes involve a specific subset of brain systems, including – but not limited to – orbitofrontal (OFC), anterior cingulate (ACC) and dorsolateral prefrontal (dlPFC) cortices [1, 2]. While the relative contribution of these subsystems is not well understood, a robust finding across studies is that orbitofrontal neurons encode value, regardless of the type of option, and whether subjects are engaged in explicit decision-making or in the subjective evaluation of single options [3–6]. Accordingly, neuropsychological studies of brain-damaged patients demonstrate that lesions to the orbitofrontal cortex induce irrational value-based decisions without impairing other types of high-level cognitive processes [7]. This means that the effective rationality of decisions hinges on the integrity of OFC circuits. But even in the absence of clear anatomical lesion, value processing in the OFC is known to exhibit systematic distortions, which can lead to irrational context-dependent behavioral biases. For example, value coding in the OFC is modulated by its pre-stimulus baseline activity [8, 9], adapts to the recent range of option values [10], and depends on whether a given option is the status-quo alternative [11] or is currently attended [12]. Taken together, these results suggest that OFC circuits are organized in such a way that they process value-related information in a moderately, yet consistently, suboptimal manner. In turn, this raises the basic question of why haven't OFC circuits evolved to minimize suboptimal distortions?

Our working assumption is that evolutionary pressure eventually selected for

OFC computations that are "rational enough", given the constraints that may act at the neurobiological level. In other words, what looks like irrational computations might actually be deemed optimal, once considering the neurobiological constraints under which brain circuits operate. A prominent example is the energetic budget of neural circuits, which encompasses both synaptic maintenance and activity-dependent firing costs [13]. These constraints are demonstrably tight: the mitochondrial metabolic supply of neurons is actively restricted at the expense of circuit-level computational efficiency [14], and a scarcity of external resources (e.g., food) eventually results in impaired neural processing [15]. This supports the idea that the brain has evolved so-called "efficient" neural coding strategies that trade off computational precision for energetic costs [16]. Interestingly, variants of such mechanisms explain value range adaptation effects in the OFC and the irrational behavioral patterns that ensue [17]. But theoretical work also emphasizes other types of tradeoffs that arise from demands on the robustness or fault-tolerance of neural circuits. A widely debated notion is that neural circuits must maintain their excitatory-inhibitory balance to ensure stability and/or homeostasis [18]. Disruption of the E/I balance has even been proposed as a core pathophysiological mechanism in several neuropsychiatric conditions [19]. Another possibility, which is pervasive in biological systems, is the need to minimize vulnerability to localized damage [20, 21]. Although direct empirical evidence for such a constraint on neural circuits is comparatively sparser, recent work indicates that neural circuits that subtend, e.g. motor behavior and working memory, achieve resilience to neural loss through architectural redundancy [22–24]. This is important because redundant neural networks are notoriously energy-inefficient [25–27]. In other words, OFC circuits may have evolved under competing architectural constraints. But then: how do we identify which neurobiological constraints might have steered OFC computations away from rationality?

We start with the premise that any constraint of the sort discussed above will

ultimately shape the architecture of OFC networks in ways that distort value computations and compromise decision rationality. This is, in fact, trivially observed in artificial neural network models of the OFC trained to perform candidate value computations while complying with these constraints (see Supplementary Material). Critically however, the form of irrational behavior that emerges depends on both the nature of the constraint and the specific value computations the OFC is assumed to perform. This is because a given type of value computation requires a tailored neural network architecture, whose native compliance with the above constraints is largely arbitrary. We thus reasoned as follows. If we knew what the OFC would look like if it had evolved in the absence of constraints, then we could compare its -counterfactual- architecture to that of actual OFC networks. We argue that artificial neural networks are valuable tools here, as their connectivity structure determines both the computations they perform and the activity patterns they exhibit in response to inputs or cues. Thus, a legitimate artificial neural network model of the OFC should exhibit activity patterns that increasingly resemble those of OFC neurons as it learns to perform the value computations that are characteristic of the OFC.

In this work, we consider the paradigmatic case of binary decisions under risk – that is, where the choice set consists of two alternatives, each defined by the probability and magnitude of prospective rewards. Numerous empirical recordings of OFC neurons are available during tasks in which macaque monkeys make such decisions. Here, we reanalyze an existing dataset in which decision cues – i.e. option-specific reward magnitude or probability – are revealed one at a time, while randomizing their sequence order across trials. This design provides a unique empirical estimate of the dynamics of information content in the OFC as value computations unfold over within-decision time [1]. In line with previous literature, we distinguish between two broad types of value computations: value *synthesis* and value *comparison*. The former implies that the OFC progressively integrates decision cues to

compute the value of both options, which can be concurrently read out on possibly orthogonal subspaces of OFC neural ensembles [5, 28–30]. The latter reduces to directly updating the value difference between the two options as a new decision cue becomes available [1, 31, 32]. Both value synthesis and value comparison can be implemented using one of five distinct neural encoding formats, which vary according to how the identity of the attended option is represented (e.g., left/right versus default/alternative), and how option values are framed (e.g., left/right, default/alternative, or attended/non-attended) [11, 33]. Together, this yields a total of ten candidate scenarios regarding OFC value computations.

We first train recursive neural networks or RNNs to perform each candidate value computation in a rational manner, given arbitrary decision cue sequences. We note that this is not a trivial task, as it requires the network to maintain a memory trace of previously attended cues, while remaining invariant to the order in which cues are presented. It turns out that RNNs can reliably learn to solve this class of problems from virtually any random initialization of their connectivity. At this point, we identify which, among these ten candidate types of value computations, yield legitimate RNN models of the OFC. To do so, we compare the full set of recorded OFC neural responses with the activity patterns of simulated RNNs exposed to the same decision trials as those experienced by the monkeys, at various stages of RNN training. As we will see, this eventually selects two specific types of value computations, which effectively are counterfactual models of OFC networks that would have evolved without any neurobiological constraint. We then distort the internal connectivity of these networks to reproduce monkeys' irrational choices in the task (about 20% of all choices). As we will show, these distorted RNNs make behavioral and neural predictions that generalize across monkeys. Finally, we compare rational and irrational RNN models of the OFC, in terms of their energetic budget, the sparsity of their connectivity structure, their E/I balance, and their robustness to neural loss. This enables us to identify which neurobiological constraint may have

shaped OFC computations.

2 Results

2.1 Identification of legitimate RNN models of OFC circuits

We took advantage of an open dataset of single unit activity recordings from the OFC, the dlPFC and the ACC of two macaque monkeys (n = 189, 135 and 183 neurons respectively) engaged in value-based decision-making (22,618 trials in total) [1, 34]. At each trial, monkeys chose between two options presented on the left and right sides of a screen, each defined by the probability and prospective amount of a rewarding juice (see Methods, Fig. 1a). Each decision cue (representing either the probability or the magnitude of the – currently attended – option) appeared sequentially and then disappeared. The monkeys could commit to a decision at any point after the second cue without necessarily sampling the remaining cues and were free to decide which cue to sample if they decided to continue the trial.

As we will see, monkeys make decisions that integrate both currently attended and remembered cues. In line with recent empirical work, we hypothesized that the OFC may implement one of two candidate decision-relevant computations: (1) computing the value of both options independently [29, 35] ("value synthesis") or (2) computing the difference between option values [5, 36] ("value comparison"). Both value synthesis and comparison can be implemented using recurrent artificial neural networks (RNNs), which operate under the same conditions as monkeys in the task. In particular, RNNs access cues sequentially and in an encoding format that specifies attribute type and rank, as well as option identity (see below). At each cue onset, these inputs are sent to a first hidden layer (cue-encoding), whose units feed their output forward to a second hidden layer (cue-integration), from which the RNN's outputs are linearly decoded (see Fig. 1b and Methods). The integration layer relies on internal recurrent connections to combine currently and

previously attended cues, and progressively update its ongoing computations [28]. Thus, value synthesis and comparison require distinct recurrent connectivity structures. Now, both value synthesis and comparison require specifying how options are identified, which is debated in the existing literature. The OFC may do so based on, e.g., spatial location [37] (left vs. right), temporal order [28] (first vs. second), or attentional focus [33] (attended vs. unattended). In principle, both OFC inputs (decision cues) and outputs (option values) may encode option identity in a different format, irrespective of whether the OFC operates value synthesis or comparison. We thus systematically tested all possible combinations, which resulted in ten cohorts of RNNs (two types of value computations combined with five input-output format variations; see Methods). Importantly, each cohort gathers a thousand RNN instances that sample the manifold of admissible connectivity structures, following random weight initializations and training datasets. Note that we did not endow RNNs with the capacity to decide which cue to attend to or when to commit to a decision; rather, we trained them to operate value synthesis or comparison independently of such processes, which are treated as arbitrary.

To begin with, we aimed to identify legitimate counterfactual, idealized RNN models of the OFC. To this end, we adopt a normative approach that obviates the need for empirical data in training RNNs. Cohorts of candidate RNNs were initialized with randomly distributed weights and subsequently trained to compute the expected value of options, as defined by rational decision theory – that is, the product of reward magnitude and probability. When tested on actual monkey decisions at the time of choice, these rational models predicted 79% of choices (monkey F: 78%, monkey M: 80%). In fact, the subjective value profiles estimated from monkey choices (see Methods) closely resemble that of expected value (see Fig. 3a and Fig. S1). Thus, rational RNNs provide a reasonable first approximation to monkeys' behavior.

Crucially, although all rational RNNs yield identical decisions in the task, their

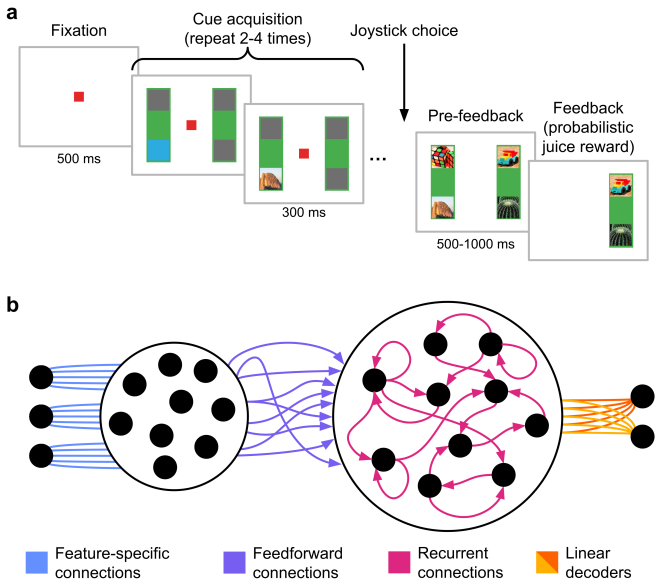

Figure 1: **Designing RNNs to solve a decision task. a,** Task design. Adapted from Hunt et al., (2018) [1]. Monkeys chose between the left and right option based on sequentially sampled informative cues representing either reward probability or magnitude. The locations of the first two cues were fixed, while subsequent cues could be freely chosen. First, a blue light indicated the location of the next available cue, which was revealed once the monkey fixated on the blue area and disappeared afterwards. The monkey could choose an option using a joystick at any point after the second cue. **b,** RNN architecture (see Methods). At each cue onset, the RNN inputs encode the currently attended cue, while the outputs are the RNN's current estimate of option values or value difference. Applying a softmax mapping to the RNN outputs yields choice probability, where options are identified with regard to spatial location, attentional focus or default status.

internal representations are different. For example, it is almost impossible to decode option values framed in a given option identity format from response patterns of RNNs that were trained under different option identity formats (see Fig. 2c). Also, individual option values are less reliably decoded from the activity of value comparison RNNs than from value synthesis RNNs (paired t-test between value synthesis and value comparison models: $p < 10^{-15}$ for all input-output format variations). We thus asked whether any of these RNN cohorts also capture key aspects of OFC neural informational geometry, despite not having been exposed to neural recordings during training. To test this, we replicated the two types of analysis conducted by Hunt et al. (2018) on single units' recordings, which we also performed on the RNNs' integration layer. We first ran a representational similarity analysis at first

cue onset, building representational dissimilarity matrices (RDMs) by correlating population activity vectors in response to all ($2 \times 2 \times 5 = 20$) possible cues (see Methods, Fig. S2 and Fig. S9). In brief, RDMs identify which cue features elicit discriminable response patterns across neurons when only a single cue is available. However, generalizing this approach to later stages of the trial becomes challenging, as RDMs face a combinatorial explosion when multiple cues have been sampled. To track neural representation geometry at all stages of decision trials, we also quantified whether and how inter-neuron differences in their sensitivity to current and past cues are preserved across cue onset times (cf. cross-correlation matrices or CCMs – see Methods). One can think of RDMs and CCMs as two distinct summary statistics of the informational geometry of distributed neural systems. We then derived the two ensuing neural distance metrics by comparing OFC neurons and RNN units at each stage of the training process (see Methods). Note that even untrained – i.e. random – RNNs exhibit some degree of neural similarity with the OFC, because they respond to value-relevant input cues. Untrained RNNs thus effectively provide the distribution of neural distances under the null. Now, when being trained to perform a specific value computation, RNNs modify their informational geometry and hence their neural distance to the OFC. We considered that legitimate RNN models of the OFC are those RNN cohorts that significantly decrease both neural distance metrics as a result of training (despite being blind to OFC activity patterns). It turns out that only two variants out of ten cohorts satisfy this selection criterion (see Fig. 2b, Fig. 3b); we only consider these for the remainder of the paper (extended results for all model variants are shown in Fig. S7 to Fig. S15 of the Supplementary Material).

In brief, both selected RNN models receive input cues that encode option identity using the temporal format, while computing option values in the attentional format. They differ only in terms of the type of value computation: one RNN cohort performs value synthesis (neural CCM distance, paired t-test: $p < 10^{-15}$, neural RDM distance, paired t-test: $p < 10^{-15}$), whereas the other performs value comparison

(CCM: $p < 10^{-5}$, RDM: $p < 10^{-15}$). Although we cannot yet arbitrate between these two scenarios, we have clearly narrowed the set of plausible counterfactual idealized OFC models.

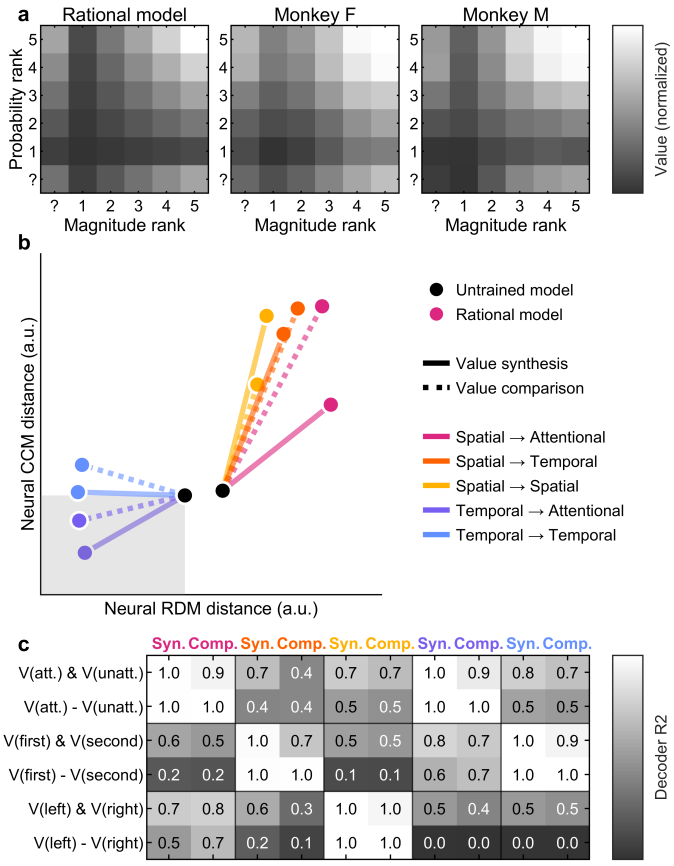

Figure 2: **Selection of candidate counterfactual idealized RNN models of the OFC. a,** Average value profiles of rational models and subjective value profiles of each monkey (fitted on choices). **b,** Neural distance trajectories between OFC and RNN cohorts during rational training. Dots show the average distance of RNN cohorts (across the 1000 RNN instances), computed using either RDMs (x-axis) or CCMs (y-axis). Black dots indicate the initial (random) state of RNN cohorts, colored dots denote their final rational state. Only two RNN cohorts significantly improve in both neural distance metrics after rational training (grey area). **c,** Information encoding in rational RNN models. Each column corresponds to a RNN cohort; each row corresponds to a type of decoded information. Numbers and grey nuances indicate the percentage of variance explained by a linear decoder applied to the RNNs' integration layer activity, averaged across the 1000 instances of the corresponding RNN cohort. All combinations are significantly better decoded than chance (paired t-test against the R2 decoded by untrained models: all $p < 10^{-15}$).

At this point, we asked whether and how counterfactual idealized OFC models need to be modified to explain irrational behavior. We thus retrained the selected

rational RNNs to predict (a subset of) monkeys' choices, of which about 20% are irrational. To preserve the interpretability of their value computations while allowing perturbations during progressive cue integration, RNNs were initialized with their trained rational weights, and retraining was restricted to recurrent connections in the integration layer. At the time of choice, retrained irrational RNNs achieved 84% choice prediction accuracy on average (monkey F: 83% (SE 1×10^{-4}), monkey M: 85% (SE 1×10^{-4})) on a test dataset, significantly outperforming rational models (paired t-test: both $p < 10^{-15}$; see Fig. 3a). Moreover, models trained on one monkey significantly outperformed their rational counterparts on the other monkey (paired t-test: both $p < 10^{-15}$; see Fig. 3a). This suggests that irrational RNNs captured hidden deterministic mechanisms underlying irrational behavior that generalize across trials and individuals.

We have leveraged the flexibility of RNNs to model both rational decision-making and systematic irrational choices, each relying on a similar structure of interconnected units. Next, we sought to determine whether irrational RNNs qualify as realistic models of OFC computations (despite not having been exposed to neural recordings during training). Remarkably, when retraining RNNs to fit the (partly) irrational behavior of monkeys, their neural distance to the OFC decreases even further compared to their rational counterparts (neural CCM distance, paired t-test; value synthesis model: $p = 9 \times 10^{-3}$, value comparison model: $p < 10^{-15}$). Furthermore, this improvement generalizes across monkeys, as shown when evaluating the neural distance of irrational RNNs to the other monkey (neural CCM distance, paired t-test: both $p < 10^{-15}$; see Fig. 3c). However, one may argue that informing RNN models about monkeys' actual choices may have facilitated the resemblance to any brain system that contributes to behavioral control in the task, thus challenging the anatomical specificity of our results. To address this point, we also computed the neural distance of irrational RNNs to dlPFC and ACC neurons. We first checked that empirical summary statistics of neural information geometry vary more across

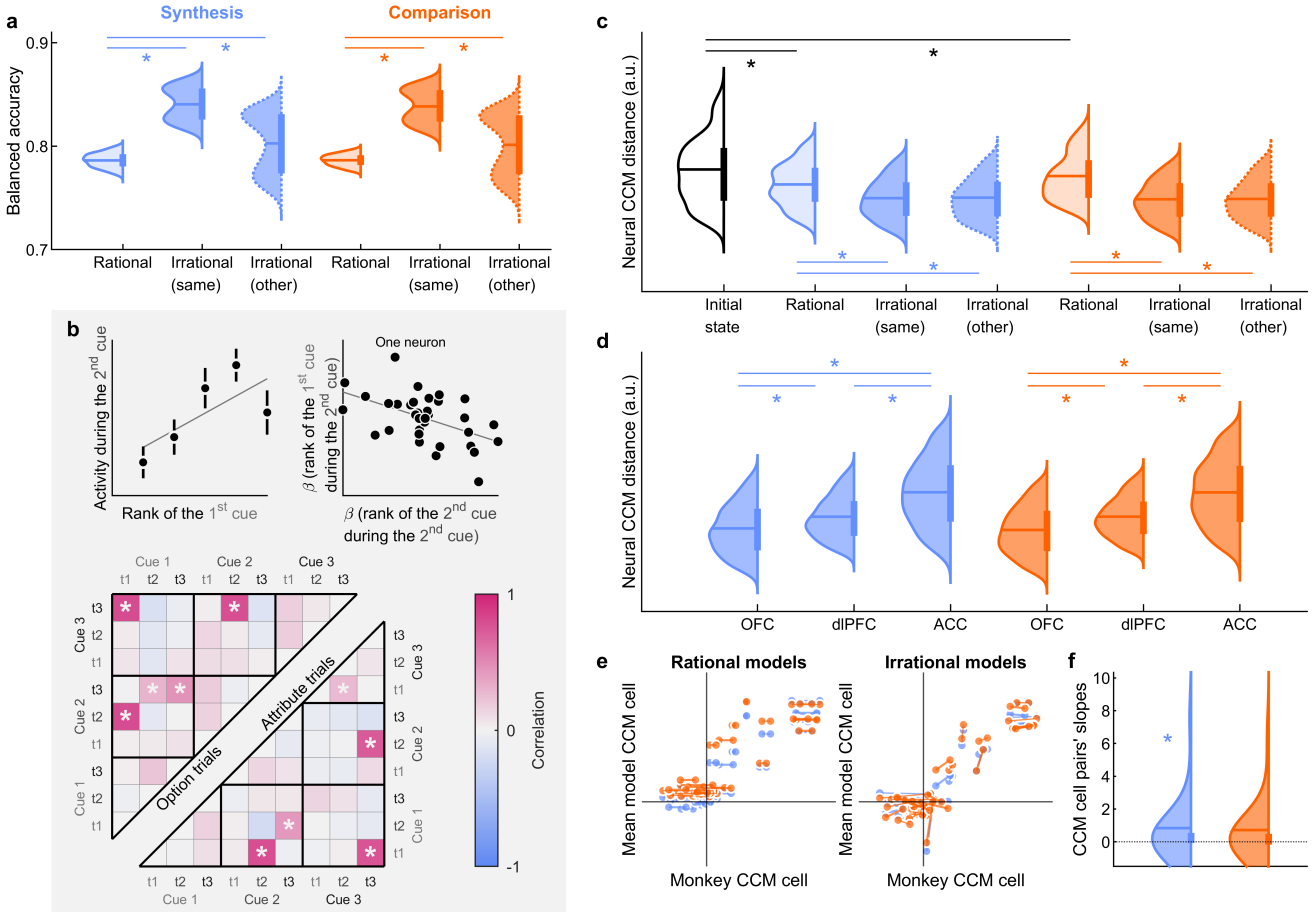

Figure 3: **Behavioral and neural realism of candidate RNN models of the OFC. a,** Balanced accuracy for predicting monkey choices. Each color corresponds to one of the two candidate models (blue: value synthesis, orange: value comparison). Lighter distributions correspond to rational models, darker distributions to irrational models, and distributions with a dashed outline represent irrational models trained on one monkey and tested on the other. Within each violin plot, the horizontal line denotes the mean, and the thicker vertical line represents the interquartile range (25th – 75th percentile). Asterisks indicate significant differences, with p-value < 0.025. **b,** Construction scheme of a CCM, applied to either OFC electrophysiological recordings or RNN activity patterns. Top left: for each OFC neuron (resp. RNN unit), mean firing rate response (resp. activity) at each cue onset is concurrently regressed across trials against the rank of all previously attended cues. Top right: correlation, across neurons (resp. units), between the ensuing regression coefficients for different cues – and possibly obtained at different onset times. Bottom: CCM: each cell in the matrix shows the correlation across neurons (resp. units) for a given pair of regression coefficients. The upper half of the matrix shows the results computed on "option trials" (where the two first cues characterize the same option), while the lower half corresponds to "attribute trials" (where the two first cues characterize the same attribute, but different options). Asterisks indicate significant correlations, with p-value < 0.001 (correction for multiple comparisons across CCM cells). **c,** Neural CCM distance between models and the OFC, same format as panel **a**. The white distribution corresponds to random RNN initializations (identical for both RNN cohorts). Asterisks indicate significant differences, with p-value < 0.0167.**d,** Neural CCM distance between irrational models and the OFC, the dlPFC and the ACC. Asterisks indicate significant differences, with p-value < 0.0167. **e,** Comparison of predicted (RNNs) and measured (OFC) CCM cells. Each color corresponds to one of the two candidate models (blue: value synthesis, orange: value comparison). Each pair of dots corresponds to a single CCM cell, for each monkey separately. Left: rational RNNs, Right: retrained (irrational) RNNs. **f,** Distribution of the slopes of CMM cell pairs in irrational RNNs (see panel **e**). Asterisks indicate significantly positive distribution, with p-value < 0.05.

brain regions than across monkeys ($p < 10^{-15}$; see Fig. S4). When comparing neural distances across brain regions, we found that irrational RNNs were significantly closer to the OFC than to the dlPFC and the ACC (neural CCM distance, paired t-test: $p < 10^{-15}$ for all comparisons between areas; see Fig. 3d).

One may also ask whether selected RNNs exhibit stereotypical trial-by-trial activity variations that are commonly observed in the OFC. First, we focused on the mixed selectivity of OFC neurons and attempted to classify units according to three distinct response profiles (see Methods): "option value cells", which encode the value of a single option (either attended or unattended); "chosen option cells", which encode the binary identity of the chosen option; and "chosen value cells", which encode the value of the chosen option (see Fig. 4a). In line with the existing literature [32, 38], we found that the trial-by-trial firing rate variations of recorded OFC neurons can be matched to one of the three response profile types at the time of choice (see Fig. 4a). Importantly, this is also the case for integration units of selected RNNs, albeit with a slight over-representation of offer value units. We also analyzed trial-by-trial variations in the grand mean activity – i.e. the average response across OFC neurons or across RNN integration units –, with the aim of verifying common fMRI findings in human OFC. In particular, we asked whether grand mean activity correlates, across trials, with either the value difference between the chosen and unchosen options (based on the monkey's choice on each trial; see Methods) or choice confidence (defined as the probability, at the time of choice, that processing the remaining unattended cues would not alter the value comparison). Consistent with previous fMRI work [3, 39], we found that the grand mean firing rate of OFC neurons significantly correlates with chosen/unchosen value difference for both monkeys (monkey F: $p = 0.048$, monkey M: $p < 10^{-10}$; see Fig. 4b) and confidence for monkey M (monkey F: $p = 0.1$, monkey M: $p < 10^{-7}$; see Fig. 4c). Interestingly, this correlation was also significantly positive, on average, in both cohorts of models, both for chosen/unchosen value difference (one-sample t-test, ra-

tional models: $p < 10^{-6}$, $p < 10^{-15}$; irrational models: $p = 5 \times 10^{-1}$, $p < 10^{-15}$; see Fig. 4b) and confidence (one-sample t-test, rational models: $p < 10^{-15}$, $p < 10^{-15}$; irrational models: $p = 1 \times 10^{-3}$, $p < 10^{-15}$; see Fig. 4c and Fig. S14).

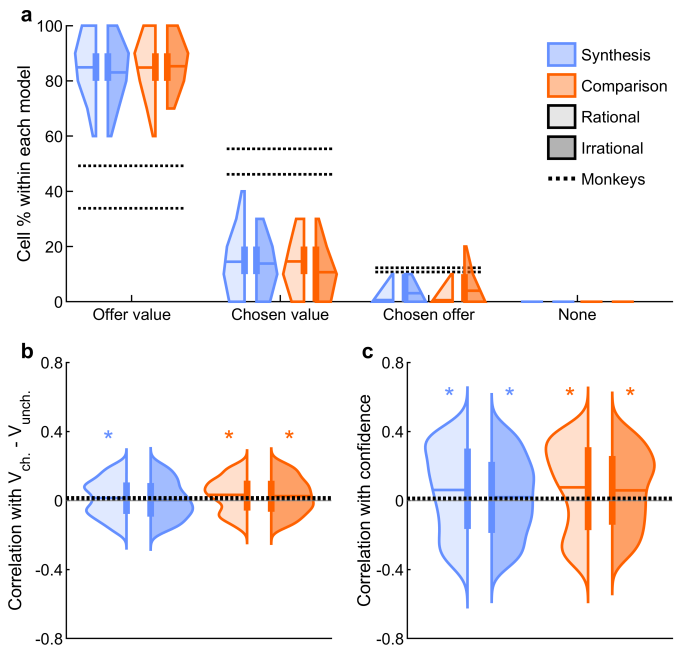

Figure 4: **Comparison of trial-by-trial activity variations between RNNs and OFC neurons. a,** Proportion of units classified as offer value, chosen value, or chosen option cells, in RNNs models and in recorded OFC neurons (at the time of choice). **b,** Correlation between the RNNs' grand mean activity and chosen/unchosen value difference. Asterisks indicate a significantly positive correlation, with p-value < 0.05. **c,** Correlation between the RNNs' grand mean activity and decision confidence. Asterisks indicate a significantly positive distribution, with p-value < 0.05.

Together, these findings suggest that the selected RNNs perform value computations that are – behaviorally and neurally – realistic. We next seek to characterize the systematic distortions in cue processing that lead to irrational choice behavior.

2.2 Analysis of computational interferences in irrational RNNs

First, we quantified potential interference effects across decision cues. Recall that, by assumption, rational choices should be solely driven by the informational content of decision cues and thus remain invariant w.r.t. cue presentation order.

In contrast, irrational interference effects would manifest as variability in RNNs' value outputs across random permutations of cue presentation order, all else being equal. We thus performed Monte-Carlo simulations of selected RNNs, quantifying the standard deviation of value outputs across randomized cue presentation orders, for all possible cue combinations and at each cue onset time (see Methods). By construction, rational RNN models exhibit almost no variability. However, irrational RNNs exhibit significantly stronger interference effects than their rational counterparts (paired t-test at each time step: both $p < 10^{-15}$). Importantly, interference effects increase as within-trial decision time unfolds (paired t-test within each cohort between step 2 and step 4: both $p < 10^{-15}$; see Fig. 5a and Fig. 5b). This suggests that systematic perturbations in sequential cue processing may accumulate over time. Accordingly, monkeys' choices become more irrational – i.e. less consistent with their average preferences – as decision time unfolds (two-sample t-test across sessions at step 2 vs. step 4, monkey F: $p < 10^{-15}$; monkey M: $p = 0.4$; at step 3 vs. step 4, monkey F: $p = 6 \times 10^{-3}$; monkey M: $p < 10^{-10}$ see Fig. 5c). One may argue that this interference effect may only be apparent, because choices that are triggered later in time may correspond to difficult decisions. Indeed, the average absolute difference between subjective option values – a proxy for decision ease – also tends to decrease when decision time increases (two-sample t-test across trials at step 2 vs. step 4, monkey F: $p < 10^{-15}$; monkey M: $p < 10^{-14}$; see Fig. 5d). To control for the effect of decision difficulty, we regressed irrational choice rates onto the absolute value difference, across trials. Reassuringly, the residuals of this regression still increase as decision time unfolds (two-sample t-test across trials, step 2 vs. step 4, monkey F: $p < 10^{-9}$; monkey M: $p = 0.03$; see Fig. S5). This means that monkeys' rationality deteriorates beyond what can be expected from decision difficulty. A possibility is that cue traces within the RNNs' integration layer may leak into one another, either across options or across attributes. To investigate this, we separated "option trials" – where the second cue reveals the missing attribute of

the same option as the first cue – from "attribute trials" – where the second cue reveals the same attribute as the first cue, but for the other option. At the second cue onset, interference effects are significantly stronger in option trials than in attribute trials, for both RNN types (paired t-test within each cohort: both $p < 10^{-15}$). This is also the case for one monkey, based on residual irrational choice rates (two-sample t-test across trials, monkey F: $p = 0.02$; monkey M: $p = 0.1$; see Fig. S5). This suggests that cue leakage effects are more pronounced within options – i.e. across attributes – than across options. Thus, we expect the integration of previously and currently attended cues to be asymmetrical, above and beyond differences induced by the type of information that they convey – i.e. reward probability vs. magnitude. To test this, we quantified the effective value output of selected RNNs as a function of the rank of both previously and currently attended cues, irrespective of cue types (see Methods). As expected, rational RNNs output values that exhibit no significant asymmetry on average (see Fig. 5e). In contrast, irrational RNNs output values that are mostly influenced by the previously attended cue (see Fig. 5f and Fig. 5f). When quantified in terms of the relative gradient of value w.r.t. the rank of previously and currently attended cues (see Methods), we find that the asymmetry is significantly stronger in irrational RNNs than in rational RNNs (paired t-test within each cohort: both $p < 10^{-15}$; see Fig. 5i and Fig. 5j). This asymmetry is also significantly present in monkeys' choices (one-sample t-test across sessions: both $p < 10^{-14}$; see Fig. 5h). These results suggest that previously attended cues leave a persisting value trace that partly resists novel value-relevant information.

In summary, irrational OFC circuits differ from their rational counterfactual variants in that they exhibit slight but systematic interference effects during value computations, which are due to peculiarities in their internal connectivity structure. We now ask whether these peculiarities may bring some form of biological advantage that may have overcompensated the behavioral irrationality that they induce.

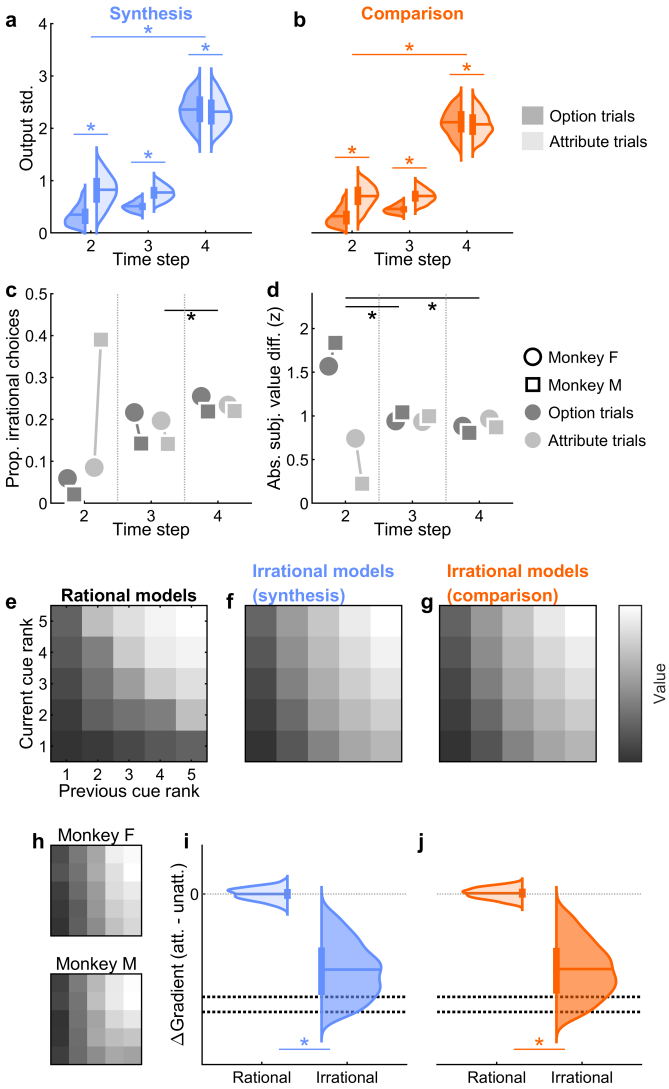

Figure 5: **Interference mechanisms in irrational models and monkeys. a,** Standard deviation of the irrational value synthesis RNNs' outputs in response to random permutations of cue sequence orders (y-axis), as a function of cue onset times (x-axis) during option trials only (light) or attribute trials only (dark). Asterisks between time steps indicate p-value < 0.05, asteris within time steps indicate p-value < 0.0167. **b,** Same format as panel **a**, but for irrational value comparison RNNs. **c,** Rate of monkeys' irrational choices (y-axis), as a function of cue onset time, for both attribute an option trials. Asterisks indicate that the difference between time steps (averaged over trial types) are significant within each monkey, with p-value < 0.0167. **d,** Absolute subjective value difference, same format as panel **c**. **e,** Average value output of rational RNNs (greyscale nuances), as a function of the rank of both previously (x-axis) and currently (y-axis) cues (see Methods). **f, g,** Same format as panel **e**, but for irrational value synthesis and value comparison RNNs, respectively. **h,** Same format as panel **e**, but for both monkeys. **i,** Average difference in the gradient of the RNNs' value output w.r.t. cue rank (attended cue minus unattended cue, see Methods), for both rational and irrational variants of value synthesis RNNs. The asterisk denotes a significant difference between rational and irrational RNNs, with p-value < 0.05. **j,** Same format as panel **i**, for value comparison RNNs.

2.3 Comparing the biological advantages of rational and irrational RNNs

First, we compared rational and irrational RNNs in terms of the metabolic cost of sustaining their respective structures. Since action potentials and synaptic maintenance are major sources of energetic consumption in the brain [40], we quantified two proxies for metabolic cost: average network activity and sparsity of their recurrent connections (see Methods). However, we found no systematic significant difference in either measure of metabolic cost between rational and irrational RNNs (paired t-test, average network activity, value synthesis: $p = 0.04$; value comparison: $p = 0.3$; connection sparsity, value synthesis: irrational less sparse than rational with $p < 10^{-11}$; value comparison: irrational more sparse than rational with $p < 10^{-15}$; see Fig. 6a and Fig. 6b).

Second, we took inspiration from other variants of efficient coding models, which rather suggests that brain circuits self-organize to maximize either information transfer rate or code sparsity. We quantify these in terms of the average log-transformed absolute gradient of units' activation function [17, 41] and the average rate of units' co-activation across all possible units pairs [42, 43], respectively (see Methods). We found no significant difference in code sparsity (paired t-test, both $p > 0.4$; see Fig. 6c). Interestingly however, we found that irrational RNNs exhibit significantly lower information transfer rate than their rational counterparts (paired t-test, both $p < 10^{-15}$; see Fig. 6d). This suggests that rational value computations may already be maximally efficient – at least w.r.t. information transfer rate. Retrospectively, this may be considered an inherent virtue of rational information processing, which precludes interference-induced information loss.

Third, we reasoned that irrational circuits may benefit from a better excitatory-inhibitory balance, which would ensure stability and/or homeostasis [18]. However, we found no significant difference in the relative proportion of negative and positive connection weights between rational and irrational RNNs (paired t-test, value

synthesis: $p = 0.2$; value comparison: $p = 0.02$; see Fig. 6e).

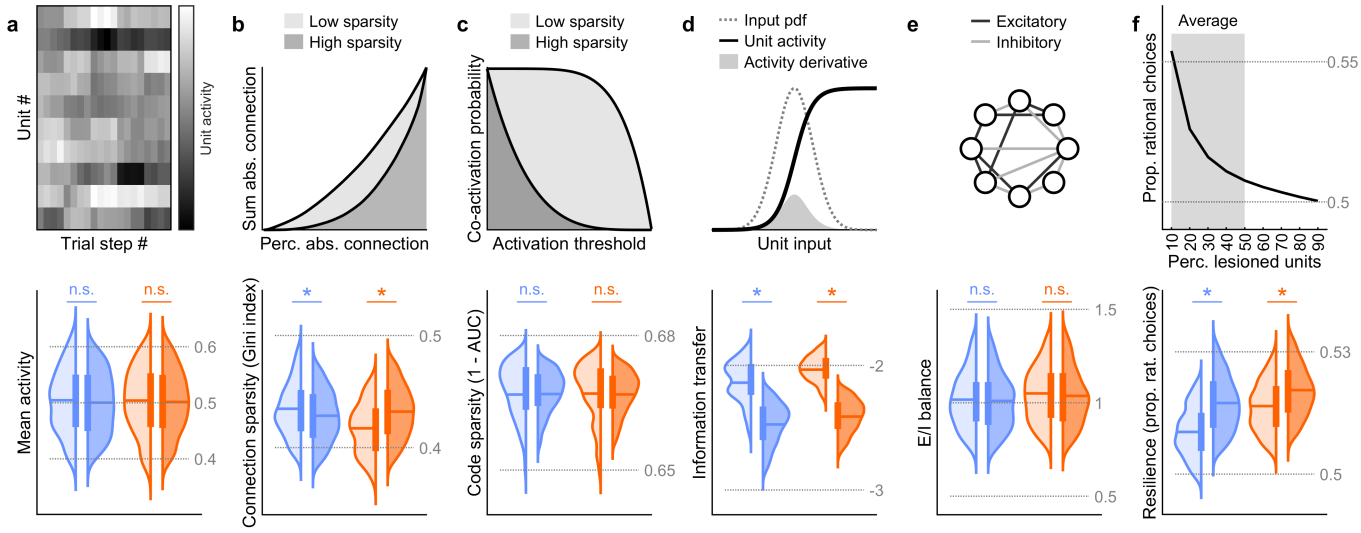

Figure 6: **Potential biological benefits of irrational circuits.** For all panels, asterisks indicate a significant difference between rational and irrational RNNs, with p-value < 0.0083 ($\frac{0.05}{6}$). Blue distributions correspond to value synthesis models, while orange distributions correspond to value comparison models. **a,** Metabolic cost, measured as the average network activity, over all trials, trial steps, and units. **b,** Code sparsity, measured as the average co-activation probability over all units pairs. **c,** Information transfer rate, measured as the average log-transformed absolute gradient of units activation function. **d,** Connection sparsity, measured using Gini index. **e,** excitatory-inhibitory balance, measured as the relative proportion of negative and positive connection weights. **f,** Tolerance to neural loss, measured as the average rational choice rate from 10% to 50% of lesioned units.

Finally, we reasoned that the internal connectivity structure of irrational circuits may enable some form of functional redundancy, which would render them more tolerant to neural loss. To test this, we simulated random virtual lesions of RNN integration units and measured the retained rate of rational choices. As expected, rational choice rate monotonically decreases when the fraction of lesioned units increases, for all types of models. Thus, we quantify neural loss tolerance to neural loss in terms of the rational choice rate averaged over lesion sizes (from 10% tp 50% of integration units, see Methods). We find that irrational RNNs exhibit significantly stronger tolerance to neural loss than their rational counterparts, irrespective of value computations (paired t-test, both $p < 10^{-15}$; see Fig. 6f).

3 Discussion

In this work, we asked whether irrational behavior may not be explained by distal constraints that act on the neurobiology of brain decision-making systems. First, we adopted a normative approach to identify idealized RNN models of the OFC, which proxy the counterfactual, unconstrained evolution of OFC circuits. We found that only a specific subset of candidate RNNs reproduces the informational geometry of the OFC – specifically, those that receive inputs encoding option identity in a temporal format (first vs. second option), while computing option values in an attentional format (attended vs. unattended option). Second, we retrained the selected RNNs to account for monkeys' irrational choices when making decisions under risk. Importantly, these retrained irrational RNNs eventually make out-of-sample behavioral and neural predictions that generalize across individuals. We also show that their peculiar internal connectivity induce deterministic interferences in value computations that explain the irrational variability of monkeys' choices across within-trial attentional trajectories. Finally, we compare the potential biological benefits of rational and irrational variants of OFC circuits and show that the latter exhibits much greater tolerance to neural loss. Irrational interferences in value computation may thus be understood as an incidental byproduct of selective pressure favoring the robustness of OFC circuits to anatomical damage.

That irrational behavior is the incidental outcome of neurobiological constraints is not a novel idea. In particular, most existing theoretical and empirical work highlight the metabolic cost of information processing in the brain [13]. To our knowledge, this work is the first attempt to demonstrate the importance of resilience to circuit damage in this context. We contend that this demonstration is theoretical in essence, at least when compared to empirical work that employ causal – e.g., genetic – manipulations to disclose proximal neurobiological constraints [14, 15]. Arguably however, it would have been difficult to provide direct empirical evidence for our main claim, at least in primates. This is inherent to the distal nature of the

constraint, which is more readily addressed from a computational perspective. In turn, our conclusions rely on a set of modeling assumptions: we will now discuss these.

To begin with, we restricted the set of candidate OFC computations to variants of value synthesis and value comparison. Although a few recent empirical studies consider other types of OFC computations [44], this prior selection is representative of current debates regarding OFC's contributions to decision making [45]. Importantly, we show that some of these variants reproduce complex features of the OFC's informational geometry, even without being informed with behavioral and/or neural data (i.e., from first principles). This includes established results regarding the mixed selectivity of OFC neural populations (cf. "option value cells", "chosen value cells" and "choice cells") [5, 38]. Moreover, we show that these computational scenarios are anatomically specific, in that their neural predictions do not resemble electrophysiological recordings in either dlPFC or ACC. Retrospectively, this assumption may thus not be so restrictive. Note that the particular RNN variants that we validated using OFC single unit recordings are consistent with landmark fMRI studies of value-based decision making. In particular, our results directly confirm fMRI studies promoting the attentional format of value coding [33]. But this is not the only possibility. For example, if a default option can be identified prior to decision onset (e.g., in terms of a prior preference over superordinate categories), then pre-stimulus activity in the OFC seems to encode its subjective value, and the strength of this response predicts peoples' irrational attachment to their default preference [11]. In other words, the OFC may use a value coding format that rather distinguishes default versus alternative options. Interestingly, this also aligns with our neural and behavioral results, under the assumption that early preferences – e.g., based upon the first attended cue – set a default option. The reason is twofold. First, as long as attention remains focused on the first option, attentional and default/alternative value-coding formats are formally indistinguishable. Second, the

persisting value trace of the firstly attended cue will, on average, appear as a bias towards the default option. In summary, although the statistical resemblance to the default/alternative hypothesis may be stronger in trials where decisions are triggered prematurely – i.e., before all relevant cues have been processed – we argue that our findings remain compatible with existing representational frameworks of value coding in the OFC. Beyond value-coding format issues, one may find it disappointing that we could not disambiguate computational scenarios of value comparison or value synthesis. The underlying question here is whether the OFC directly implements choice, or whether its role is limited to assigning values to available options [28, 46]. When implemented in the form of winner-take-all networks, the former scenario explains reproduced findings in electrophysiological and neuroimaging studies, in particular: the observed mixed selectivity of OFC cells [5, 32], as well as the apparent encoding of the value difference between chosen and unchosen options – at least during late phases of decision making [47]. Interestingly, we have shown that such findings can be equally well reproduced by RNNs performing either value synthesis or value comparison. This calls for experiments that are designed to distinguish these kinds of computational scenarios, as opposed to testing one of them.

Also, we did not vary the global architecture of our artificial neural nets, which consisted of a layer of feature-encoding units sending their outputs to a layer of recurrently connected integration units. In line with recent neural net approaches to value computations in the OFC [17, 28], we adopted the minimal architecture that ensures universal approximation capabilities while using a limited number of sigmoidal units [48, 49]. Note that a major computational bottleneck of both value synthesis and value comparison scenarios is OFC circuits' capacity for combining value-relevant attributes of arbitrary number and type [35]. Now, the above two-layer architecture provides a flexible and simple solution to this problem that rests on the second layer's trained ability to integrate arbitrary sequences of attributes, whose type and rank are encoded in separate pools of the first layer units. In

particular, this circumvents the need for otherwise unrealistic, context-dependent changes in connectivity with upstream brain systems involved in recognizing or storing value-relevant information. Nevertheless, the relative simplicity of our design contrasts with previous studies that favored off-the-shelf deep neural nets to approximate the hierarchical organization of, e.g., primates' visual ventral stream [50] or humans' language networks [51]. From a machine learning perspective, tasks such as visual perception and speech comprehension are inherently difficult problems, which remained unsolved until the advent of deep neural networks trained on massive labeled datasets. In these domains, objective task performance reliably predicts statistical similarity with neural data. This relationship, however, does not generalize to our findings: RNNs tend to more closely resemble OFC data when they permit systematic, error-inducing interferences. In retrospect, it is remarkable that our value synthesis/comparison RNNs exhibit such realistic features, at both the behavioral and neural levels. This is despite the degeneracy of RNN wiring profiles w.r.t. each type of value computation, which we systematically explored by repeating the training process across many random initializations of RNN parameters. Arguably, the ensuing marginalization process renders our results robust to local minima issues. This statistical benefit would have been prohibitively costly to match using deep neural net architectures.

One might also argue that rational and irrational RNNs may have been compared in an unfair manner. For example, we chose to train rational RNNs under a normative approach, which precludes idiosyncratic variations in risk attitudes. The rationale here was to obtain neural nets that could serve as neutral and fully interpretable reference points, in that their computational objective was under our control – i.e. computing expected values, as prescribed by decision theory. We acknowledge that, when it comes to measuring statistical similarity to neural recordings, irrational RNNs may somehow benefit from being trained on individual behavioral datasets. However, the fact that irrational RNNs make out-of-sample predictions

that generalize across individuals rather suggests that they have captured hidden, yet shared, decision mechanisms. In any case, there is no reason to think that this training difference would, in principle, favor irrational RNNs in terms of resilience to circuit damage. A related concern is whether the latter may be the artefactual byproduct of re-training, which may – in principle – provide an additional opportunity for improving efficiency or robustness. This is the reason why we also explored another training strategy for irrational RNNs, which starts from the same randomly initialized parameter sets as rational RNNs. As evidenced in the Results section (see also Fig. S7, Fig. S8 and Fig. S15), our conclusions remain unchanged under this alternative training strategy.

In conclusion, we believe our modeling assumptions are tenable, at least when compared to state-of-the-art computational studies in the field. They enabled us to reverse the usual approach to disclosing distal neurobiological constraints on rationality, which typically rests on highlighting conflicts with the demands of behavioral performance (cf. Fig. S6). In contrast, we identify realistic mechanisms that explain observed deviations to rationality, and explore their potential neurobiological advantages. We believe that this may be a fruitful method for investigating related evolutionary or developmental issues in cognitive neuroscience.

4 Methods

4.1 Task design

Monkeys were seated in a behavioral chair with their heads restrained. Each trial began when the monkey fixated on a central fixation cue for 500 ms. At the start of the trial, two options were presented, each consisting of two hidden cues initially masked by grey squares. One of these squares then turned blue, indicating the first cue available for sampling. When the subject fixated on the blue square, the corresponding picture cue was revealed and had to be continuously fixated for

300 ms before it was re-masked.

All picture cues had been previously learned and were associated with either probability or magnitude information. Probability cues indicated reward probabilities of 10%, 30%, 50%, 70%, or 90%, while magnitude cues represented reward magnitudes of 0.15, 0.35, 0.55, 0.75, or 0.95 arbitrary units (AU).

Following the initial cue, a second blue square highlighted the next available cue, which had to be sampled using the same procedure. This second cue was either the other cue of the same option (option trial) or the cue of the other option associated with the same attribute (attribute trial). After the second cue, the two remaining cues were simultaneously highlighted with blue squares, allowing the subject to freely choose which one to sample next, or to select one of the two options using a joystick. If a third cue was sampled, the subject could then either sample the final cue or make a choice. Once the fourth cue was revealed, the subject was required to make a choice.

4.2 Neural data

The designing of the task, behavioral and neural datacollection were entirely performed by Hunt et al. 2018 [1], and published in an open dataset [34].

Neuronal activity was recorded from three brain regions in each monkey: the orbitofrontal cortex (OFC), the anterior cingulate cortex (ACC) and the dorsolateral prefrontal cortex (dlPFC). During each session, neurons were simultaneously recorded from two or all three regions using between 8 and 24 electrodes. Neurons with a firing rate below 1 Hz were excluded. In total, for monkey F, 108 neurons were retained in the OFC, 97 in the ACC, and 107 in the dlPFC. For monkey M, 87 neuron were retained in the OFC, 49 in the dlPFC, and 101 in the ACC. These recordings were collected across 24 session for monkey F and 29 sessions for monkey M. Within each subject and brain area, neurons were pooled into pseudopopulations on which all subsequent analyses were performed.

To enable direct comparison with RNN models, which operate in discrete time, we averaged each neuron's firing rate over a 100-400 ms window following cue onset. This provided a single activity measure per neuron per trial time step, consistent with the temporal granularity of activity in the RNNs.

4.3 Value profile estimation

We estimated the subjective value profile of each monkey (and each model) using standard statistical procedures, based solely on the agent's choices. More precisely, we fitted the underlying value function, under the assumption that choices followed a simple softmax mapping of the difference in option values:

$$p(\text{choose option 1}) = \frac{1}{1 + \exp\left(-\left(V(p_1, m_1) - V(p_2, m_2)\right)\right)} \tag{1}$$

where p_i and m_i denote the reward probability and magnitude of option i, as known by the agent at the time of choice, and $V(p, m)$ is the corresponding subjective value. Equation (1) provides a binomial likelihood function for observed choices, given the unknown monkeys' value function. Parameterizing the value function then enables us to regress trial-by-trial choices against option attributes. To allow for maximal modelling flexibility, we employed a semi-parametric approach, whereby each possible combination of probability and magnitude – including cases in which one or both attributes were unknown at the time of choice – is captured using a specific model parameter. In other words, the only modelling constraint here is that the same value function applies to all options, but its functional form remains unconstrained.

4.4 RNN architecture

Let $t \in \{1, 2, 3, 4\}$ denote the time step index at which cue is revealed or attended within a decision trial. The RNN component variables are defined as follows:

- $\overrightarrow{x}(t) \in \mathbb{R}^3$: Inputs vector at time t. These include the attribute rank and type – probability or magnitude –, as well as the identity of the currently attended option (see below).
- $\overrightarrow{L_1}(t) \in \mathbb{R}^9$: Unit activation vector in the first hidden layer at time t.
- $\overrightarrow{L_2}(t) \in \mathbb{R}^{10}$: Unit activation vector in the second hidden layer at time t.
- $\overrightarrow{y}(t) \in \mathbb{R}^1$ (for value comparison models) or $\overrightarrow{y}(t) \in \mathbb{R}^2$ (for value synthesis models): Output prediction at time t.

At the first time step ($t = 1$), information propagates through the network according to the following equations:

$$\overrightarrow{L_1}(t) = f\left(W_{\text{encode}} \cdot \overrightarrow{x}(t) - \overrightarrow{b_1}\right) \tag{2}$$

$$\overrightarrow{L_2}(t) = f\left(W_{\text{forward}} \cdot \overrightarrow{L_1}(t) - \overrightarrow{b_2}\right) \tag{3}$$

$$\overrightarrow{y}(t) = W_{\text{readou}} \cdot \overrightarrow{L_2}(t) \tag{4}$$

At later time steps ($t > 1$), the second hidden layer incorporates recurrent activity elicited by the previous cues. This means that Equation (4) is replaced with:

$$\overrightarrow{L_2}(t) = f\left(W_{\text{forward}} \cdot \overrightarrow{L_1}(t) + W_{\text{recurrent}} \cdot \overrightarrow{L_2}(t-1) - \overrightarrow{b_2}\right) \tag{5}$$

Here, $W_{\blacksquare}$ refers to matrices of connection weights, and $\overrightarrow{b_{\blacksquare}}$ are bias vectors applied to the corresponding hidden layers. The weights W_{encode} and biases $\overrightarrow{b_1}$ where initially set such that each admissible cue rank (x_1) preferentially activated a dedicated unit in a rank-specific pool of first layer units. Similarly, each admissible cue type (x_2) and option identity (x_3) preferentially activated one out of two units each (again in secluded pools of first layer units). To ensure distributed encoding within each pool, the activation profiles of first layer units were configured to tile the domain of their

specific input uniformly: whenever one unit's activity reached 75% of its maximum, the "adjacent" units in the pool were 25% active.

To impose a biologically plausible constraint on firing rates, we used a sigmoid activation function f for all units in the hidden layers:

$$f : x \mapsto \frac{1}{1 + \exp(-x)} \tag{6}$$

Importantly, when structurally organized into two hidden layers, neural nets with a limited number of sigmoidal units possess universal approximation capabilities [48, 49].

The RNN received inputs one at a time, in a sequential manner – as monkeys did in the task. The sequence order is determined by the exogenous control of attention, which samples cues in an arbitrary fashion within a decision trial. Let $x_1(t)$, $x_2(t)$ and $x_3(t)$ denote the components of the input vector $\overrightarrow{x}(t) \in \mathbb{R}^3$:

- $x_1(t)$ encodes the normalized rank of the attended cue, with the following mapping:

Magnitude cue	Probability cue	Cue rank	x_1
0.15 AU	10%	1	0.1
0.35 AU	30%	2	0.3
0.55 AU	50%	3	0.5
0.75 AU	70%	4	0.7
0.95 AU	90%	5	0.9

- $x_2(t)$ encodes the attribute type. Probability: $x_2 = 0$; magnitude: $x_2 = 1$.
- $x_3(t)$ encodes the identity of the attended option. Option 1: $x_3 = 0$; option 2: $x_3 = 1$.

Note that the identity of the attended option can be expressed in two different representation formats: spatial (left vs. right) or temporal (first vs. second). This distinction affects the encoding of x_3, as illustrated in the following example trials:

Trial ID	Attended option side	$x3$ in the *spatial* frame (right = 0, left = 1)	x_3 in the *temporal* frame (first = 0, second = 1)
1	Right	0	0
1	Left	1	1
1	Left	1	1
1	Right	0	0
2	Left	1	0
2	Left	1	0
2	Left	0	1
2	Left	0	1

Similarly, the outputs of the network can be expressed in different representation formats: spatial, temporal, or attentional (attended vs. unattended). The example trials below illustrate how the encoding format of option values varies across these frames. Let V_{left} and V_{right} denote the values of the left and right options as estimated by the network at each cue onset. The statistical similarity between representation formats depend on the actual sequence order of cue attendance:

Trial ID	Attended option side	Output in the *spatial* frame	Output in the *temporal* frame	Output in the *attentional* frame
1	Right	V_{right} & V_{left}	V_{right} & V_{left}	V_{right} & V_{left}
1	Left	V_{right} & V_{left}	V_{right} & V_{left}	V_{left} & V_{right}
1	Left	V_{right} & V_{left}	V_{right} & V_{left}	V_{left} & V_{right}
1	Right	V_{right} & V_{left}	V_{right} & V_{left}	V_{right} & V_{left}
2	Left	V_{right} & V_{left}	V_{left} & V_{right}	V_{left} & V_{right}
2	Left	V_{right} & V_{left}	V_{left} & V_{right}	V_{left} & V_{right}
2	Right	V_{right} & V_{left}	V_{left} & V_{right}	V_{right} & V_{left}
2	Right	V_{right} & V_{left}	V_{left} & V_{right}	V_{right} & V_{left}

Note that not all combinations of input/output formats are trainable. More precisely, when the input's option identity is encoded using the spatial format, then

value outputs can be encoded in all representation formats (3 possibilities). However, when the input's option identity is encoded using the temporal format, then the spatial information is lost, which leaves only 2 possible value encoding formats (temporal and attention frames). This means that there is only 5 combinations of input/output representation formats in total.

4.5 RNN training

4.5.1 Rational training

Models were implemented and trained using MATLAB R2022b with the VBA toolbox [52]. The RNN parameters subject to training (W_{forward}, $W_{\text{recurrent}}$ and $\overrightarrow{b_2}$) were initialized as samples from an i.i.d. Gaussian distribution with mean 0 and variance 0.5. For each RNN model, the training procedure was repeated with a different initial random sample, until 1000 trained models reached 95% test accuracy. In the main text, we refer to the ensemble of trained RNNs as a "cohort", each of which corresponds to a given type of value computation (value synthesis versus value comparison) and a given combination of input/output representation format (see above).

For each model instance in a given RNN cohort, a training set and a testing set consisting of 500 trials each were generated. Every trial consisted of a sequence of four cues, randomly chosen among the set of different option pairings, and presented in a random order. Note that training and testing trials could be classified post-hoc as either "attribute trials" or "option trials", depending on whether the attention switched to the second option at the second cue onset, or not.

Now, so-called "value synthesis" models were trained to output the expected value of both options in response to each cue presentation. In contrast, "value comparison" models were trained to output the difference in expected value between the two options. When both the probability and magnitude of an option were

available, its expected value was computed as their product. If any attribute was missing, its rank was replaced by its prior mean under the task distribution.

Training was terminated when the absolute change in variational free energy between VBA successive iterations fell below 10. A network was considered successfully trained if it reached at least 95% of explained variance on its testing set. Each RNN cohort consisted of 1000 independently trained model instances, each with a unique training set, testing set, and parameter initialization. Importantly, random seeds were shared across cohorts, which allowed for fully matched comparisons across cohorts.

4.5.2 Irrational training

To preserve the interpretability of value computation and input/output representation formats, all network parameters were frozen except for $W_{\text{recurrent}}$. The network outputs were transformed into choice probabilities via a simple softmax mapping: $p(\text{choose option 1}) = \frac{1}{1+\exp(-\Delta V)}$.

In contrast to the rational training phase, where value outputs were evaluated at each cue onset, irrational training evaluated the value outputs only at the time of choice. Since $w_{\text{recurrent}}$ controls the way RNNs assimilate cue sequences to perform their specific value computations, this effectively restricts the admissible sources of irrational behavior to within-trial interferences between cues.

Each RNN instance within each cohort was then re-trained to fit the choices of each individual monkey, using a training dataset of 2000 trials randomly selected from monkeys' recorded sessions. This procedure produces two twin versions of each retrained irrational model – one for each monkey. We then test their respective behavioral and neural predictions within and across monkeys. The former evaluates their inter-trial generalization ability, whereas the latter focuses on inter-individual generalization ability.

In a supplementary analysis, we also trained networks to predict monkey choices

directly from their initial parameterization, without a prior rational training phase. This procedure was thus similar to rational training in terms of training load (cf. optimization of all parameters in VBA and no partial freezing of parameters), except that value outputs were only evaluated at the time of choice.

4.5.3 Rational training with constraints

In another supplementary analysis, we trained RNNs to perform rational value computations while simultaneously satisfying neurobiological constraints. More precisely, RNN parameters were trained to optimize a tradeoff between the accuracy of their value outputs and the compliance to one of the following constraints: minimal average firing rate, maximal connection sparsity (considering both feedforward and recurrent weights), maximal coding efficiency, or maximal resilience to neural loss (see *Biological benefits* below). To balance these two – possibly conflicting – objectives, we introduced trade-off weights that varied logarithmically from 10^{-3} to 10^{3}, allowing us to modulate the relative importance of "behavioral efficiency" (accuracy of value outputs) versus "neural efficiency" (compliance to the neural constraint). The results of this training procedure can be eyeballed in Fig. S6.

4.6 Analysis of informational geometry within neural populations: summary statistics

4.6.1 Representational similarity analysis

Let $\overrightarrow{L_2^x}(1)$ denote the vector of activations in the RNN's second layer in response to input $\overrightarrow{x}$ at the first cue onset. This vector can be computed for each possible input $\overrightarrow{x_k}$, which yields 20 distinct activation patterns (i.e., 5 cue ranks $\times 2$ cue types $\times 2$ options). The representational dissimilarity matrix (RDM) is constructed element by element by computing pairwise similarities between these activation vectors [53]:

$$RDM_{k,l} = r\left(\overrightarrow{L_2^{x_k}}(1), \overrightarrow{L_2^{x_l}}(1)\right) \tag{7}$$

where r denotes Pearson's correlation. If $RDM_{k,l}$ strongly positive, then activity patterns are mostly invariant to differences between inputs $\overrightarrow{x_k}$ a,d $\overrightarrow{x_l}$, i.e. the neural representation of these inputs are similar. In brief, RDMs enables us to identify what input features need to change to elicit distinct neural responses.

The same procedure is applied to recordings of OFC neurons (as well as to neural recordings within the dlPFC and the ACC), using vectors of averaged firing rates measured between 100 ms and 400 ms following the first cue onset. This yields two RDMs: one for the model (RDM^{model}) and one for the OFC data (RDM^{OFC}). Full RDM summary statistics for all monkeys and brain regions can be eyeballed in Fig. S2, and average RDMs obtained for all RNN cohorts are plotted in Fig. S9.

Finally, the similarity between these matrices is quantified using a rank-based distance metric:

$$\text{dist}_{\text{RDM}} = 1 - \rho\left(RDM^{\text{OFC}}_{\text{upper}}, RDM^{\text{model}}_{\text{upper}}\right) \tag{8}$$

Here, ρ denotes Spearman's correlation and RDM_{upper} refers to the upper triangular half of the matrix, excluding the diagonal. We used a rank-based metric because experimental neural data is typically much noisier than model activations, resulting in compressed correlation ranges that are more appropriately captured by rank correlations. The neural RDM distance trajectories between all models and brain areas can be eyeballed in Fig. S7, and the details of the comparison with OFC recordings are displayed in Fig. S13.

4.6.2 Cross-correlation matrices

Unfortunately, the above representational similarity analysis does not scale well with the number of input combinations. In our context, its statistical cost is pro-

hibitive for later phases of decision trials, when more than one cue has been attended. For example, at the second cue onset, there are 400 possible cue combinations, which would induce RDMs with almost 79800 entries. This is why we resort to another type of summary statistics, which was proposed by Hunt et al. (2018) [1]. In brief, this analysis enables us to quantify and compare the multiple traces that cue sequences leave on units' activity, at the cost of partly neglecting differences induced by attribute types. This simplifying assumption exploits the observed quasi-symmetrical impact of reward probability and magnitude on monkeys' subjective value profiles (see Fig. 2a).

Let $L_2^{s(x)}(i,t)$ denote the activation of unit i in the second hidden layer after the presentation of a cue at time $t \in \{1,2,3\}$, given a sequence of inputs $s(x)$ of length t. We regress each second layer unit's trial-by-trial activity variations at cue onset t concurrently onto trial-by-trial variations of normalized attribute rank in all cues, while identifying cues by their appearance order in the sequence. Note that we also include two additional regressors, which encode how consistent the 2nd and 3rd cues (respectively) are w.r.t. the currently preferred option, as well as an intercept term. This approach aims at detecting nontrivial memory traces of previously attended cues, while ruling out mere confirmation effects in value coding neurons. Importantly, we separate "option trials" (where the first two cues belong to the same option) from "attribute trials" (where the first two cues describe the same attribute – i.e. probability or magnitude – but for both options) prior to performing the regression analyses. This yields one set of regression coefficient estimates per trial type.

Let $\overrightarrow{\beta_k}(t) \in \mathbb{R}^{n_{\text{units}}}$ denote the vector of t-statistics associated with regression coefficient estimates for the kth attended cue ($k \in \{1,2,3\}$), given each second layer unit's activity at time t. This vector measures how sensitive to the kth attended cue second layer units are (at time t) in normalized signal-to-noise ratio units. This enables a direct quantitative comparison across units, cue presentation orders and

decision times. Note that $\overrightarrow{\beta_k}(t)$ vectors that involve cue presentation orders that are strictly higher than activity sampling times (i.e. when $k > t$) are statistically meaningless.

We then define the cross-correlation matrix (CCM) as follows:

$$CCM_{k,k',t,t'} = \rho\left(\overrightarrow{\beta_k}(t), \overrightarrow{\beta_{k'}}(t')\right) \tag{9}$$

where ρ denotes Pearson's correlation. A strongly positive CCM cell indicates that the neurons most sensitive to the kth attended cue at time t arealso those most sensitive to the k'^{th} cue at time t'.

We obtain full CCMs by systematically varying cue presentation orders (k and k') as well as activity sampling times (t and t'), yileding a 9 by 9 symmetrical matrix. We then remove CCM cells that are meaningless to avoid statistical illusions possibly induced by imperfections in trial randomizations. We repeat this process for both trial types (cf. "option trials "versus "attribute trials"), yielding two CCM types. Differences between the two types of CCM cells that involve the first and second cue onset times (i.e. $CCM_{1,2,\blacksquare,\blacksquare}$) signal that a shift in the attended option affects the network's distributed computations. In particular, if neurons respond to the value difference between options, then one expects $CCM_{1,2,2,2}$ to be positive for option trials, and negative for attribute trials [1].

We apply the same analysis on recorded data from OFC neurons (as well as neurons in the dlPFC and ACC). For each neuron, we compute the average firing rate in a 100-400 ms window after each cue onset and regress it against normalized attribute ranks of all cues (including the same additional regressors). This provides summary statistics whose temporal resolution matches that of RNN models. Full CCM summary statistics for all monkeys and brain regions can be eyeballed in Fig. S3, average CCMs obtained for all RNN cohorts are plotted in Fig. S9 and the distribution of key CCM cells are shown in Fig. S10.

To compare the informational geometry of RNNs and OFC neural populations,

we simply compute the Euclidian distance between the meaningful CCM cells:

$$\text{dist}_{\text{CCM}} = \left\| \begin{bmatrix} \text{vec}(CCM^{\text{OFC}}_{\text{option}}) \\ \text{vec}(CCM^{\text{OFC}}_{\text{attribute}}) \end{bmatrix} - \begin{bmatrix} \text{vec}(CCM^{\text{model}}_{\text{option}}) \\ \text{vec}(CCM^{\text{model}}_{\text{attribute}}) \end{bmatrix} \right\|_2 \tag{10}$$

The neural CCM distance trajectories between all models and brain areas can be eyeballed in Fig. S7 and Fig. S8, and the details of the comparison with OFC recordings are displayed in Fig. S13.

4.6.3 Mixed selectivity: offer value cells, chosen value cells and choice cells

To identify offer value, chosen value, and choice cells, we replicated the analysis previously introduced by Padoa-Schioppa and colleagues [38]. When applied to neural recordings in the OFC, we relied on subjective value profiles, as estimated from monkeys' choices in the task (see *Value profile estimation*). To maximize the match between analyses, we also use model-specific value profiles for RNNs.

For each unit, we performed four separate regressions across all trials, using four distinct regressors: the value of option 1, the value of option 2, the value of the chosen option, and the identity of the chosen option. Note that we match the option identity encoding format to the one used by each RNN model. Each unit was assigned to the category that yielded the highest percentage of explained variance, provided the regression was significant (p-value < 0.05). Otherwise, no category was assigned. The distribution of cell categories for all models can be eyeballed in Fig. S14.

4.7 Analysis of computational interferences in irrational RNNs

4.7.1 Dependency on cue sequence order

In principle, rational behavior in the task only depends upon the content of value-relevant information, but not on its presentation sequence order. Under this view, any observed dependency on cue sequence order violates rationality.

Let $y^{s(x)}(t)$ denote the value difference between options, as can be readout from the RNN's response to an input sequence $s(x)$ of length t – where the sequence $s(x)$ is composed of a series of cues presented in a specific order. For value synthesis models, we compute $y^{s(x)}(t)$ by subtracting the readouts of both option values (at time t). To quantify the dependency on cue presentation order, we first measure the standard deviation of $y^{s(x)}(t)$ across all possible permutations of cue orderings while keeping the set of t attended cues constant, and then average the results over cue sets. We repeat this process separately for option trials and attribute trials, meaning that we only consider cue order permutations that are admissible for each trial type.

Let X be the set of all possible combinations of t cues, and for each such set $x \in X$, let $S(x)$ denote the set of admissible orderings of those cues (restricted to the relevant trial type). Then, the model's dependency on sequence order at time t, denoted $d(t)$, is defined as:

$$d(t) - \frac{1}{|X|} \sum_{x \in X} \sqrt{\mathrm{Var}\left(\{y^{s(x)}(t) | s \in S(x)\}\right)} \tag{11}$$

Note that this measure is defined for all decision times starting from the second cue onset ($t \geq 2$) – and both trial types. This enables us to track the possible accumulation of interferences in RNN computations as decision time unfolds.

Models' dependency on sequence order is represented in Fig. S12 (top row) for

all cohorts.

Note that this analysis cannot be directly applied to monkeys' choices, as we cannot have access to the monkeys' internal value estimates for each cue sequence order. This is because the total number of unique cue sequence orders is very large: specifically, 10000 per trial type (corresponding to 5 cue ranks for each of the 4 cues and $4! = 24$ possible cue orderings, restricted to valid ones). This number is comparable to the total number of decision trials for each monkey (Monkey F: 9463 trials; Monkey M: 13155 trials), which means that we have no empirical repetitions of cue sequence orders. This is the reason why we resort to measures of apparent deviations to rational choice, which effectively reduce to detecting trials that are incongruent with estimates of monkeys' subjective preferences (see Fig. 5c and Fig. 5d).

4.7.2 Persisting value traces

The above dependency on sequence order may be partly driven by a directional bias, whereby the effective weight of each cue is determined by its onset time. For example, previously attended cues may weigh more on value outputs than currently attended cues, all else being equal. We developed a specific method for detecting such persisting value traces, which can be equally applied to both RNN simulations and monkeys' behavior in the task.

We start by re-estimating value profiles, while allowing for value differences between options that are currently or previously attended (at the time of choice), and having separated trials by the type of attended cue (reward probability vs magnitude). Let $V_{\text{att}}^{\text{prob}}$ denote the pseudo-value function of the attended option when a probability cue is attended at the time of choice, and $V_{\text{unatt}}^{\text{prob}}$ that of the other (unattended) option. Let p_{att} and m_{att} be the ranks of the attended option's probability and magnitude, and p_{unatt} and m_{unatt} those of the unattended option. The choice probability for selecting the attended option is given by:

$$p(\text{choose attended option}) = \frac{1}{1 + \exp\left(-\left(V_{\text{att}}^{\text{prob}}(p_{\text{att}}, m_{\text{att}}) - V_{\text{unatt}}^{\text{prob}}(p_{\text{unatt}}, m_{\text{unatt}})\right)\right)} \tag{12}$$

This provides a binomial likelihood function for observed choices that are triggered when a probability cue is attended. To estimate the pseudo-value profiles $V_{\text{att}}^{\text{prob}}$ and $V_{\text{unatt}}^{\text{prob}}$, we use the same semi-parametric approach as before. The pseudo-value profiles $V_{\text{att}}^{\text{mag}}$ and $V_{\text{unatt}}^{\text{mag}}$ can be estimated similarly, given observed choices that are triggered when a magnitude cue is attended.

Recall that $V_{\text{att}}^{\text{prob}}$ (resp. $V_{\text{att}}^{\text{mag}}$) is the pseudo-value that ensues from currently attending a probability (resp., a magnitude) cue, while the magnitude (resp., probability) cue was previously attended (if ever). To quantify the relative impact of currently and previously attended cues while marginalizing over cue types, we then combine $V_{\text{att}}^{\text{prob}}$ and $V_{\text{att}}^{\text{mag}}$ to form the following average pseudo-value profile V_{att}:

$$V_{\text{att}} = \frac{1}{2}\left(V_{\text{att}}^{\text{prob}} + V_{\text{att}}^{\text{mag}\top}\right) \tag{13}$$

Importantly, Vatt is a 6 by 6 pseudo-value profile whose first dimension (columns) spans the rank of the currently attended cue, while its second dimension (rows) spans the rank of the previously attended cue – including the case where it is unknown at the time of choice. A rational agent would exhibit a strictly symmetric average pseudo-value profile.

To quantify potential asymmetries in V_{att}, we computed gradients of V_{att} with respect to the currently and previously attended (or, equivalently, unattended) dimensions. Let $V_{\text{att}}(:, i)$ denote the yth row (i.e., fixed attended attribute, varying unattended attribute) and $V_{\text{att}}(i, :)$ denote the ith column (i.e., fixed unattended attribute, varying attended attribute). Average pseudo-value gradients are given by:

$$\begin{cases} \frac{\partial V\text{att}}{\partial \text{att}} & = \frac{1}{5\times 4}\sum_{i=1}^{5}\sum_{j=1}^{4} V_{\text{att}}(i,j+1) - V_{\text{att}}(i,j) \\ \frac{\partial V\text{att}}{\partial \text{unatt}} & = \frac{1}{5\times 4}\sum_{i=1}^{4}\sum_{j=1}^{5} V_{\text{att}}(i+1,j) - V_{\text{att}}(i,j) \end{cases} \tag{14}$$

These gradients capture the average rate of change in the average pseudo-value profile w.r.t. changes in the attended or unattended attribute ranks. For example, a stronger gradient along the unattended dimension signals a greater sensitivity to the previously attended cue. This is the hallmark of a persisting value trace that resists novel (currently attended) information. Results can be eyeballed for all RNN moels in Fig. S12.

4.8 Biological benefits

4.8.1 Efficient coding: average network firing rate

The average network firing rate $\bar{f}$ of a model is defined as the average activation of RNNs' second layer units, across all units, time steps, and possible trials:

$$\bar{f} = \frac{1}{N_S \times N_t \times N_i} \sum_{s(x)\in S(X)} \sum_{t=1}^{N_t}\sum_{i=1}^{N_i} L_2^{s(x)}(i,t) \tag{15}$$

where $S(X)$ denotes the set of all admissible sequences of 4 cues, $N_S = 10000$ is the number of such sequences, $N_t = 4$ is the number of cues per trial, and $N_i = 10$ is the number of units in the RNNs' second hidden layer.

This is a proxy for the network's metabolic or energetic consumption.

4.8.2 Efficient coding: code sparsity

We quantify the sparsity of activations in the second hidden layer based on the statistical overlap of unit activations across trials. Specifically, we define code sparsity as a decreasing function of the likelihood of multiple units being simultaneously active, relative to their typical activity distributions.

Let us say that unit i is "active" if its response $L_2(i)$ strictly exceeds the ath

percentile of its marginal activity distribution, where $a \in [0, 100]$ is an arbitrary activation threshold (expressed in the normalized units of cumulative distributions). Let $N_{\text{active}}(a, s(x), t)$ denote the number of active units at decision time t, for the input sequence $s(x)$, under the threshold a. The probability that two randomly selected units are simultaneously active is computed as:

$$P(a, s(x), t) = \frac{N_{\text{active}}(a, s(x), t)\,(N_{\text{active}}(a, s(x), t) - 1)}{N_i(N_i - 1)} \tag{16}$$

Finally, the code sparsity S is defined as:

$$S = 1 - \frac{1}{101 \times N_S \times N_t} \sum_{a=0}^{100} \sum_{s(x) \in S(X)} \sum_{t=1}^{N_t} P(a, s(x), t) \tag{17}$$

When S tends towards unity, code sparsity is maximal, i.e. units almost never co-activate across trials and decision time steps.

4.8.3 Efficient coding: information transfer rate

For a given network unit, information transfer rate is maximal when the noise-induced information loss is minimal, i.e. when the entropy of the unit's output (across sampled cue sequences) is maximal. Let $f : x \mapsto y$ be the input-output activation function of neural net units. At the low noise limit, information transfer rate IR is defined as the expected, log-transformed, absolute gradient of the activation function [41]:

$$IR = \mathbb{E}\left[\ln\left|\frac{\partial f}{\partial x}(x)\right|\right] \tag{18}$$

Here, each RNN's second layer unit i receives a linear combination of activations from the first hidden layer and recurrent activations from itself at previous time steps, which are passed through a sigmoid activation function (with bias):

$$f(x) = \frac{1}{1 + \exp(-x + b)} \tag{19}$$

The derivative of the sigmoid simplifies to:

$$\frac{\partial f}{\partial x}(x) = f(x)\,(1 - f(x)) \tag{20}$$

Therefore, the network's average information transfer rate reduces to:

$$AIR = \frac{1}{N_S \times N_t \times N_i} \sum_{s(x)\in S(X)} \sum_{t=1}^{N_t} \sum_{i=1}^{N_i} \ln\left(L_2^{s(x)}(i,t) \times \left(1 - L_2^{s(x)}(i,t)\right)\right) \tag{21}$$

where $L_2^{s(x)}(i,t)$ denotes the activation of unit i at step t in response to the input sequence $s(x)$.

4.8.4 Connection sparsity

We quantify the sparsity of RNNs' recurrent connections using the Gini index [54], computed over the absolute values of the entries $(w_i)_{i\in\{1,\cdots n\}}$ in the recurrent weight matrix $W_{\text{recurrent}}$. The weights are first sorted in ascending order of their absolute magnitude, such that $|w_1| \leq |w_2| \leq \cdots |w_n|$. The Gini index reflects the degree of unequal sharing of connection strengths across all pairs of connected units:

$$G = 1 - \frac{2}{n\sum_{i=1}^{n}|w_i|} \sum_{i=1}^{n} |w_i| \left(n - i + \frac{1}{2}\right) \tag{22}$$

A Gini index close to 1 indicates high sparsity, which proxies a low synaptic maintenance cost. Note that fault-tolerance is typically achieved using high functional redundancy (i.e. low sparsity), though this is not a necessary condition.

4.8.5 E/I balance

The excitatory/inhibitory balance of a circuit refers to the relative contribution of excitatory and inhibitory inputs on features of the circuit's evoked responses (e.g., selective tuning). In electrophysiological studies, E/I balance is usually eval-

uated using intracellular conductance estimates across a wide range of conditions and contexts. Here, we quantify a structural E/I balance, which we define as the ratio between the number of positive and strictly negative connection weights. This measure includes all hidden-layer connections, encompassing both the feedforward weights W_{forward} and the recurrent weights $W_{\text{recurrent}}$. Formally:

$$E/I \text{ balance} = \frac{\#\left\{w \geq 0 | w \in W_{\text{forward}} \bigcup W_{\text{recurrent}}\right\}}{\#\left\{w < 0 | w \in W_{\text{forward}} \bigcup W_{\text{recurrent}}\right\}} \tag{23}$$

Note that RNNs that exhibit mostly excitatory connections (E/I balance $\gg 1$) may exhibit divergent activity dynamics, which precludes accurate value computations (at least in late phases of decision trials).

4.8.6 Resilience to neural loss

Let $n \in \{0, 1, ..., N_i\}$ denote the number of lesioned units in the second hidden layer, and let $C_n \in \{1, ..., N_i\}^n$ be a combination of such n units. Lesioning a unit was done by externally setting its activation to 0 across all time steps and trials. Let $z_{model}(s(x), t, C_n) \in \{0, 1\}$ denote the RNN's simulated choice in response to an input sequence $s(x)$ at time t, under a lesion C_n of its integration layer. Let $z_{\text{rational}}(s(x), t)$ denote the rational choice (i.e. the preferred option based upon options' expected value) for the same input sequence and time step. We define the resilience to neural loss R_{rational} as the retained rational choice rate, averaged over all possible lesion configurations involving 10% to 50% of all units in the second hidden layer:

$$R_{\text{rational}} = \frac{1}{5 \times N_S \times N_t} \sum_{n=1}^{5} \frac{1}{\binom{10}{n}} \sum_{C_n \in C(n)} \sum_{s(x) \in S(X)} \sum_{t=1}^{N_t} 1_{\{z_{\text{model}}(s(x),t,C_n) = z\text{rational}(s(x),t)\}} \tag{24}$$

where $C(n)$ denotes the set of possible combinations of n units within an ensemble of 10 units. When R_{rational} tends towards unity, the behavioral outputs of RNNs are

unaffected by virtual lesions.

We also computed an alternative metric, $R_{\text{consistent}}$, by comparing the lesioned model's behavior to the choice of its own non-lesioned counterpart (which may deviate from rational expected values):

$$R_{\text{consistent}} = \frac{1}{5 \times N_S \times N_t} \sum_{n=1}^{5} \frac{1}{\binom{10}{n}} \sum_{C_n \in C(n)} \sum_{s(x) \in S(X)} \sum_{t=1}^{N_t} 1_{\{z_{\text{model}}(s(x),t,C_n) = z\text{model}(s(x),t,C_0)\}} \tag{25}$$

Resilience to circuits' damage can also be evaluated using virtual lesions of connections within the network. In this analysis, a proportion $n \in \{10, 20, 30, 40, 50\}$ % of the RNN's connection weights are set to 0,and resilience to neural loss is measured as the retained rational choice rate. Note that we did this separately for recurrent connections only ($W_{\text{recurrent}}$) and for all hidden-layer connections ($W_{\text{forward}} \bigcup W_{\text{recurrent}}$). All results can be eyeballed on Fig. S15.

References

[1] Laurence T. Hunt et al. "Triple dissociation of attention and decision computations across prefrontal cortex". In: *Nature Neuroscience* 21.10 (2018), pp. 1471–1481. DOI: 10.1038/s41593-018-0239-5.

[2] Matthew F. S. Rushworth et al. "Frontal Cortex and Reward-Guided Learning and Decision-Making". In: *Neuron* 70.6 (2011), pp. 1054–1069. DOI: 10.1016/j.neuron.2011.05.014.

[3] Alizée Lopez-Persem et al. "Four core properties of the human brain valuation system demonstrated in intracranial signals". In: *Nature neuroscience* 23.5 (2020), pp. 664–675. DOI: 10.1038/S41593-020-0615-9.

[4] Maël Lebreton et al. “An automatic valuation system in the human brain: evidence from functional neuroimaging”. In: *Neuron* 64.3 (2009), pp. 431–439. DOI: 10.1016/j.neuron.2009.09.040.

[5] Camillo Padoa-Schioppa and Katherine E. Conen. “Orbitofrontal Cortex: A Neural Circuit for Economic Decisions”. In: *Neuron* 96.4 (2017), pp. 736–754. DOI: 10.1016/J.NEURON.2017.09.031.

[6] Oscar Bartra, Joseph T. McGuire, and Joseph W. Kable. “The valuation system: A coordinate-based meta-analysis of BOLD fMRI experiments examining neural correlates of subjective value”. In: *NeuroImage* 76 (2013), pp. 412–427. DOI: 10.1016/J.NEUROIMAGE.2013.02.063.

[7] Lesley K. Fellows. “Deciding how to decide: ventromedial frontal lobe damage affects information acquisition in multi-attribute decision making”. In: *Brain: A Journal of Neurology* 129.Pt 4 (2006), pp. 944–952. DOI: 10.1093/brain/awl017.

[8] Raphaëlle Abitbol et al. “Neural Mechanisms Underlying Contextual Dependency of Subjective Values: Converging Evidence from Monkeys and Humans”. In: *Journal of Neuroscience* 35.5 (2015), pp. 2308–2320. DOI: 10.1523/JNEUROSCI.1878-14.2015.

[9] Camillo Padoa-Schioppa. “Neuronal Origins of Choice Variability in Economic Decisions”. In: *Neuron* 80.5 (2013), pp. 1322–1336. DOI: 10/gg39j4.

[10] Camillo Padoa-Schioppa. “Range-Adapting Representation of Economic Value in the Orbitofrontal Cortex”. In: *Journal of Neuroscience* 29.44 (2009), pp. 14004–14014. DOI: 10.1523/JNEUROSCI.3751-09.2009.

[11] Alizée Lopez-Persem, Philippe Domenech, and Mathias Pessiglione. “How prior preferences determine decision-making frames and biases in the human brain”. In: *eLife* 5 (2016). Ed. by Michael J Frank, e20317. DOI: 10.7554/eLife.20317.

[12] Antonio Rangel, Colin Camerer, and P. Read Montague. "A framework for studying the neurobiology of value-based decision making". In: *Nature Reviews. Neuroscience* 9.7 (2008), pp. 545–556. DOI: 10.1038/nrn2357.

[13] Zahid Padamsey and Nathalie L. Rochefort. "Paying the brain's energy bill". In: *Current Opinion in Neurobiology* 78 (2023), p. 102668. DOI: 10.1016/j.conb.2022.102668.

[14] Anjali Amrapali Vishwanath et al. *Mitochondrial Ca2+ efflux controls neuronal metabolism and long-term memory across species*. 2024. DOI: 10.1101/2024.02.01.578153.

[15] Zahid Padamsey et al. "Neocortex saves energy by reducing coding precision during food scarcity". In: *Neuron* 110.2 (2022), 280–296.e10. DOI: 10.1016/j.neuron.2021.10.024.

[16] D. Attwell and S. B. Laughlin. "An energy budget for signaling in the grey matter of the brain". In: *Journal of Cerebral Blood Flow and Metabolism: Official Journal of the International Society of Cerebral Blood Flow and Metabolism* 21.10 (2001), pp. 1133–1145. DOI: 10.1097/00004647-200110000-00001.

[17] Jules Brochard and Jean Daunizeau. "Efficient value synthesis in the orbitofrontal cortex explains how loss aversion adapts to the ranges of gain and loss prospects". In: *eLife* (2024). DOI: 1.

[18] Lu Chen et al. "Homeostatic plasticity and excitation-inhibition balance: The good, the bad, and the ugly". In: *Current Opinion in Neurobiology* 75 (2022), p. 102553. DOI: 10.1016/j.conb.2022.102553.

[19] Vikaas S. Sohal and John L. R. Rubenstein. "Excitation-inhibition balance as a framework for investigating mechanisms in neuropsychiatric disorders". In: *Molecular Psychiatry* 24.9 (2019), pp. 1248–1257. DOI: 10.1038/s41380-019-0426-0.

[20] Saket Navlakha et al. "Topological properties of robust biological and computational networks". In: *Journal of The Royal Society Interface* 11.96 (2014), p. 20140283. DOI: 10.1098/rsif.2014.0283.

[21] Hiroaki Kitano. "Towards a theory of biological robustness". In: *Molecular Systems Biology* 3.1 (2007), p. 137. DOI: 10.1038/msb4100179.

[22] Guang Chen et al. "Modularity and robustness of frontal cortical networks". In: *Cell* 184.14 (2021), 3717–3730.e24. DOI: 10.1016/j.cell.2021.05.026.

[23] Shyam Srinivasan and Charles F. Stevens. "Robustness and fault tolerance make brains harder to study". In: *BMC Biology* 9.1 (2011), pp. 1–3. DOI: 10.1186/1741-7007-9-46.

[24] Muneki Ikeda et al. *Circuit Degeneracy Facilitates Robustness and Flexibility of Navigation Behavior in C.elegans*. 2018. DOI: 10.1101/385468.

[25] Matthew Chalk, Olivier Marre, and Gašper Tkačik. "Toward a unified theory of efficient, predictive, and sparse coding". In: *Proceedings of the National Academy of Sciences* 115.1 (2018), pp. 186–191. DOI: 10.1073/pnas.1711114115.

[26] W. B. Levy and R. A. Baxter. "Energy efficient neural codes". In: *Neural Computation* 8.3 (1996), pp. 531–543. DOI: 10.1162/neco.1996.8.3.531.

[27] Torsten Hoefler et al. "Sparsity in deep learning: pruning and growth for efficient inference and training in neural networks". In: *The Journal of Machine Learning Research* 22.1 (2021), 241:10882–241:11005.

[28] Mathias Pessiglione and Jean Daunizeau. "Bridging across functional models: The OFC as a value-making neural network". In: *Behavioral neuroscience* 135.2 (2021), pp. 277–290. DOI: 10.1037/BNE0000464.

[29] Vincent B. McGinty and Shira M. Lupkin. "Behavioral read-out from population value signals in primate orbitofrontal cortex". In: *Nature Neuroscience* 26.12 (2023), pp. 2203–2212. DOI: 10/gs78nc.

[30] Justin M. Fine et al. “Abstract Value Encoding in Neural Populations But Not Single Neurons”. In: *Journal of Neuroscience* 43.25 (2023), pp. 4650–4663. DOI: 10.1523/JNEUROSCI.1954-22.2023.

[31] Caleb E. Strait, Tommy C. Blanchard, and Benjamin Y. Hayden. “Reward Value Comparison via Mutual Inhibition in Ventromedial Prefrontal Cortex”. In: *Neuron* 82.6 (2014), pp. 1357–1366. DOI: 10/f56xkp.

[32] Sébastien Ballesta and Camillo Padoa-Schioppa. “Economic Decisions through Circuit Inhibition”. In: *Current Biology* 29.22 (2019), 3814–3824.e5. DOI: 10.1016/j.cub.2019.09.027.

[33] Seung-Lark Lim, John P. O’Doherty, and Antonio Rangel. “The Decision Value Computations in the vmPFC and Striatum Use a Relative Value Code That is Guided by Visual Attention”. In: *Journal of Neuroscience* 31.37 (2011), pp. 13214–13223. DOI: 10.1523/JNEUROSCI.1246-11.2011.

[34] Laurence T. Hunt, W.M. Nishantha Malalasekera, and Steven W. Kennerley. *Recordings from three subregions of macaque prefrontal cortex during an information search and choice task*. CRCNS.org. DOI: 10.6080/K0PZ5712. URL: https://crcns.org/data-sets/pfc/pfc-7.

[35] John P. O’Doherty, Ueli Rutishauser, and Kiyohito Iigaya. “The hierarchical construction of value”. In: *Current opinion in behavioral sciences* 41 (2021), p. 71. DOI: 10.1016/J.COBEHA.2021.03.027.

[36] Matthew FS Rushworth et al. “Valuation and decision-making in frontal cortex: one or many serial or parallel systems?” In: *Current Opinion in Neurobiology*. Decision making 22.6 (2012), pp. 946–955. DOI: 10/f4hkkh.

[37] Camillo Padoa-Schioppa and John A. Assad. “The representation of economic value in the orbitofrontal cortex is invariant for changes of menu”. In: *Nature neuroscience* 11.1 (2008), pp. 95–102. DOI: 10.1038/NN2020.

[38] Camillo Padoa-Schioppa and John A. Assad. "Neurons in the orbitofrontal cortex encode economic value". In: *Nature* 441.7090 (2006), pp. 223–226. DOI: 10.1038/NATURE04676.

[39] Erie D. Boorman et al. "How Green Is the Grass on the Other Side? Frontopolar Cortex and the Evidence in Favor of Alternative Courses of Action". In: *Neuron* 62.5 (2009), pp. 733–743. DOI: 10.1016/j.neuron.2009.05.014.

[40] Sharna D. Jamadar et al. "The metabolic costs of cognition". In: *Trends in Cognitive Sciences* 0.0 (2025). DOI: 10.1016/j.tics.2024.11.010.

[41] Jean-Pierre Nadal and Nestor Parga. "Nonlinear neurons in the low-noise limit: a factorial code maximizes information transfer". In: *Network: Computation in Neural Systems* 5.4 (1994), pp. 565–581. DOI: 10.1088/0954-898X_5_4_008.

[42] Dong-Ping Yang, Hai-Jun Zhou, and Changsong Zhou. "Co-emergence of multi-scale cortical activities of irregular firing, oscillations and avalanches achieves cost-efficient information capacity". In: *PLoS computational biology* 13.2 (2017), e1005384. DOI: 10.1371/journal.pcbi.1005384.

[43] Junya Hirokawa et al. "Frontal cortex neuron types categorically encode single decision variables". In: *Nature* 576.7787 (2019), pp. 446–451. DOI: 10.1038/s41586-019-1816-9.

[44] Nir Moneta, Shany Grossman, and Nicolas W. Schuck. "Representational spaces in orbitofrontal and ventromedial prefrontal cortex: task states, values, and beyond". In: *Trends in Neurosciences* (2024). DOI: 10.1016/j.tins.2024.10.005.

[45] Eric B. Knudsen and Joni D. Wallis. "Taking stock of value in the orbitofrontal cortex". In: *Nature Reviews. Neuroscience* 23.7 (2022), pp. 428–438. DOI: 10.1038/s41583-022-00589-2.

[46] Thelma Landron et al. “Dissociation of Value and Confidence Signals in the Orbitofrontal Cortex during Decision-Making: An Intracerebral Electrophysiology Study in Humans”. In: *Journal of Neuroscience* 45.18 (2025). DOI: 10.1523/JNEUROSCI.1740-24.2025.

[47] Laurence T. Hunt et al. “Mechanisms underlying cortical activity during value-guided choice”. In: *Nature neuroscience* 15.3 (2012), pp. 470–476. DOI: 10.1038/NN.3017.

[48] Vitaly Maiorov and Allan Pinkus. “Lower bounds for approximation by MLP neural networks”. In: *Neurocomputing* 25.1 (1999), pp. 81–91. DOI: 10.1016/S0925-2312(98)00111-8.

[49] Namig J. Guliyev and Vugar E. Ismailov. “Approximation capability of two hidden layer feedforward neural networks with fixed weights”. In: *Neurocomputing* 316 (2018), pp. 262–269. DOI: 10.1016/j.neucom.2018.07.075.

[50] Daniel L. K. Yamins et al. “Performance-optimized hierarchical models predict neural responses in higher visual cortex”. In: *Proceedings of the National Academy of Sciences of the United States of America* 111.23 (2014), pp. 8619–8624. DOI: 10.1073/PNAS.1403112111.

[51] Charlotte Caucheteux, Alexandre Gramfort, and Jean-Rémi King. “Evidence of a predictive coding hierarchy in the human brain listening to speech”. In: *Nature Human Behaviour* 7.3 (2023), pp. 430–441. DOI: 10.1038/s41562-022-01516-2.

[52] Jean Daunizeau, Vincent Adam, and Lionel Rigoux. “VBA: A Probabilistic Treatment of Nonlinear Models for Neurobiological and Behavioural Data”. In: *PLOS Computational Biology* 10.1 (2014), e1003441. DOI: 10.1371/journal.pcbi.1003441.

[53] Nikolaus Kriegeskorte. “Relating population-code representations between man, monkey, and computational models”. In: *Frontiers in Neuroscience* 3 (2009). DOI: 10.3389/neuro.01.035.2009.

[54] Trevor Gale, Erich Elsen, and Sara Hooker. *The State of Sparsity in Deep Neural Networks*. 2019. arXiv: 1902.09574[cs,stat].