Value Representations in the Rodent Orbitofrontal Cortex Drive Learning, not Choice

Kevin J. Miller[1,2,3], Matthew M. Botvinick[2,4,6], and Carlos D. Brody[1,5,6]

[1] Princeton Neuroscience Institute, Princeton University, Princeton, NJ, USA
[2] DeepMind, London, UK
[3] Department of Ophthalmology, University College London, London, UK
[4] Gatsby Computational Neuroscience Unit, University College London, London, UK
[5] Howard Hughes Medical Institute and Department of Molecular Biology, Princeton University, Princeton NJ, USA
[6] Co-corresponding authors

Abstract

Humans and animals make predictions about the rewards they expect to receive in different situations. In formal models of behavior, these predictions are known as value representations, and they play two very different roles. Firstly, they drive *choice*: the expected values of available options are compared to one another, and the best option is selected. Secondly, they support *learning*: expected values are compared to rewards actually received, and future expectations are updated accordingly. Whether these different functions are mediated by different neural representations remains an open question. Here we employ a recently-developed multi-step task for rats that computationally separates learning from choosing. We investigate the role of value representations in the rodent orbitofrontal cortex, a key structure for value-based cognition. Electrophysiological recordings and optogenetic perturbations indicate that these representations do not directly drive choice. Instead, they signal expected reward information to a learning process elsewhere in the brain that updates choice mechanisms.

Introduction

Representations of expected value play a key role in human and animal cognition[1–3]. A key brain region involved in representing and using expected value information is the orbitofrontal cortex (OFC)[4–7]. Theoretical accounts of the OFC variously emphasize roles for this region in either choosing[8,9] or learning[10–12]. Lesion studies in primates have suggested that different subregions of OFC may selectively support learning and choosing[13–15]. In rodents, considerable evidence suggests that disrupting neural activity in the OFC impairs learning[16–20], but it remains controversial whether this also impairs choice[21–23]. Recording studies in many species have revealed neural correlates of expected value in the OFC[24–28], but it remains unclear whether these representations are selective for roles in learning and in choosing.

Studies reporting neural representations of expected value in the OFC have typically adopted a repeated-trial experimental design. In each trial, the subject is presented with one or more options, chooses among them, and receives a reward associated with the chosen option. Some studies[27–30] present options that are novel or have repeatedly-changing reward associations in order to elicit trial-by-trial learning, interleaving decision-making with learning by design. Other studies[26,31–35] present options whose reward associations are stable and well-learned, with the intention of isolating decision-making specifically. Although not incentivized by task design, trial-by-trial learning is frequently evident in these tasks as well[35–37]. The fact that learning and choosing are intertwined in these tasks, operating at similar times and over the same set of items, has made it difficult to conclusively assign a role for OFC value representations in one or the other process.

Previous studies designed to separate learning from choosing have typically used a different set of behavioral methods. These experiments adopt a temporal strategy, attempting to behaviorally separate the time of learning from the time of choosing, and performing neural experiments specifically at one time or the other. They have provided compelling evidence that the rodent OFC is involved in learning: silencing the OFC specifically during the learning phase of either of two associative learning procedures (blocking or overexpectation) impairs subsequent behavior on an assay performed with the OFC intact[17,19]. Similarly, recordings of the OFC during the learning phase of a related procedure (sensory preconditioning) reveal evidence of expected outcome coding[20]. Experiments of this type have also been interpreted as indicating that the OFC is involved in choosing: silencing the OFC specifically during the post-learning assay of either of two procedures (sensory preconditioning or outcome devaluation) impairs expression of knowledge gained earlier[14,19,21,38].

One limitation of methods that attempt to separate learning and choosing temporally is that these are fundamentally cognitive events, internal to the brain, and their timing cannot be fully controlled by an experimenter. For example, it has been suggested that subjects faced with a choice might reconsider information that was presented in the past; cognitively, this might best be interpreted as a learning process happening during a putative choice epoch[39–41]. Conversely, subjects experiencing a surprising outcome might consider its implications for future choices that they expect to be faced with, and memorize decisions for later use[42]; cognitively, this might best be interpreted as a choice process happening during a putative learning epoch. A second limitation is that these studies typically address forms of learning that unfold over many sessions, a very different timescale and behavioral regime to the one in which OFC value representations are typically studied. Since different neural mechanisms may be at play in trial-by-trial vs. session-by-session learning, it is difficult to confidently interpret the results of neural recording experiments addressing the former in light of neural silencing experiments addressing the latter.

Here, we develop an alternative approach to separating learning from choosing, which is based not only on *when* they happen, but also on *what content* they operate over. The opportunity to do this arises from rat behavior in a multi-stage value-based task[43] adapted from the human literature (the "two-step" task[44]). In previous work[43] we demonstrated that this task elicits a particular behavioral strategy ("model-based planning"), and that whole-session silencing of the OFC was sufficient to impair this strategy. Those results were consistent with the idea that the

OFC supports planning, and value-guided cognition in general, but did not distinguish whether the OFC plays a role in learning, in choosing, or both. Here, we take advantage of the fact that the planning strategy separates task items into some that are chosen between (but not learned about), and some that are learned about (but not chosen between; **Fig. 1c**). A computational model, fit to the ongoing behavior, infers trial-by-trial estimates[45] of the values of these two separate types of items. We can thus ask how OFC neuron firing rates are related to each of the two types of items. We find that neurons in the OFC correlate with the model's estimates of the expected values of the items being learned about, but only weakly with the expected values of the items being chosen between. This indicates that neural representations in the OFC carry value information that is selective for a role in learning, but do not seem to carry value information selective for a role in choice.

Using the fact that our task allows separation of the concurrent learning and choosing processes, we then asked whether OFC silencing would selectively impair one or the other process. We found that transiently silencing OFC activity on a subset of trials, specifically during the time period when representations of values learned about were the strongest, was sufficient to impair behavior. Furthermore, the detailed pattern of this impairment was reproduced in our computational model if we disrupted the role of expected value in learning, but not the role of expected value in choice. In this model, learning still happens even when the OFC is silenced, but it happens in an impaired way due to one of its key input signals, expected value, being disrupted. Taken together with the literature, these results suggest that, even when choosing and learning are commingled on the same trial-by-trial timescale, value representations in the rodent OFC do not drive choices directly, but instead support a learning process that updates choice mechanisms elsewhere in the brain.

Results

Two-Step Task Separates Roles for Choosing, Learning

We trained rats to perform a multistep decision-making task[43] that separates the computational functions of learning and choosing. In this task, a choice that the rat makes on the first step is linked probabilistically to an outcome that occurs on the second step, which in turn is linked probabilistically to reward (**Fig. 1a**). The first-step transition probabilities are fixed for each rat, while the second-step reward probabilities change unpredictably and must be continually learned about. This task structure, in which choices and outcomes are linked probabilistically rather than having a 1-to-1 relationship, combined with the particular strategy rats use to solve it, provides the critical separation necessary to differentiate learning from choosing. Specifically, on the first step of every trial, rats choose an option, based on values that are computed, not learned ("Use World Model" and "Make Choice" in **Fig. 1c**); while on the second step they learn the values of the outcomes ("Update Outcome Values" in **Fig. 1c**) but are led to those outcomes without making a choice.

The structure of each trial was as follows: The rat first initiated the trial by poking its nose into a neutral center port, and then selected one of two choice ports (**Fig. 1a i,ii**). One choice caused a left outcome port to become available with high probability ("common" transition), and a right outcome port to become available with low probability ("uncommon" transition), while the opposite choice reversed these probabilities (**Fig. 1a iii)**. Following the initial choice, an auditory

cue informed the rat which of the two outcome ports had in fact become available on that trial and, after poking into a second neutral center port, the available outcome port was further indicated by a light (**Fig. 1a iv,v**). The rat was required to poke into the available outcome port (no choice in this step), where it received a water reward with a given probability (**Fig. 1a v,vi**). The two outcome ports differed in the probability with which they delivered reward, and these reward probabilities changed at unpredictable intervals (**Fig. 1b**). The subjects were thus required to continually update their estimates of the expected reward probability at each outcome port *o* (which we label $V(o)$), through a learning process in which they compare their expectations to actual reward received ("Update Outcome Values" in **Fig. 1c**).

Previously[43], we have shown that rats solve the task using a particular strategy termed "model-based planning"[46–48]. This strategy utilizes an internal model of action-outcome relationships to compute the expected values of the two choice ports (which we label $Q(c)$). A planning strategy results in very different computational roles for outcome-port values and choice-port values. Outcome port values, (V), are learned incrementally from recent rewards (**Fig. 1c**, "Learn Outcome Port Values"). Choice port values (Q) are computed based on the outcome port values and the known transition probabilities between first-step choice and second-step outcome (the "world model", $P(o|c)$; **Fig. 1c**, "Compute Choice Port Values"). These choice port values are then used to determine the next choice (**Fig. 1c**, "Choose a Choice Port"). The values of the choice ports therefore drive choice directly, while the learned values of the outcome ports support choice only indirectly, by directly supporting learning. Consistent with previous results[43], rats in the current study adopted such a planning strategy, as indicated both by an index quantifying the extent to which the rats' behavior is sensitive to the world model (the "planning index" **Fig. 1d**, see Methods, **Fig S1**, **Fig S2**), and by fits of an artificial planning agent (**Fig. 1e**). This artificial agent matches rat behavior on the task, and allows us to probe the role of the OFC in two key ways. First, it can provide trial-by-trial estimates of the value placed by the rat on the choice ports and on the outcome ports. We will use these estimates[45], to ask whether each type of value information is reflected in the activity of OFC neurons. Second, the agent can be altered to selectively disrupt information about expected values of the choice ports (Q) or of the outcome ports (V) on particular trials, and used to generate synthetic behavioral datasets that predict the consequences of these specific impairments. We compare these to behavioral data from rats performing the task while undergoing optogenetic silencing of the OFC.

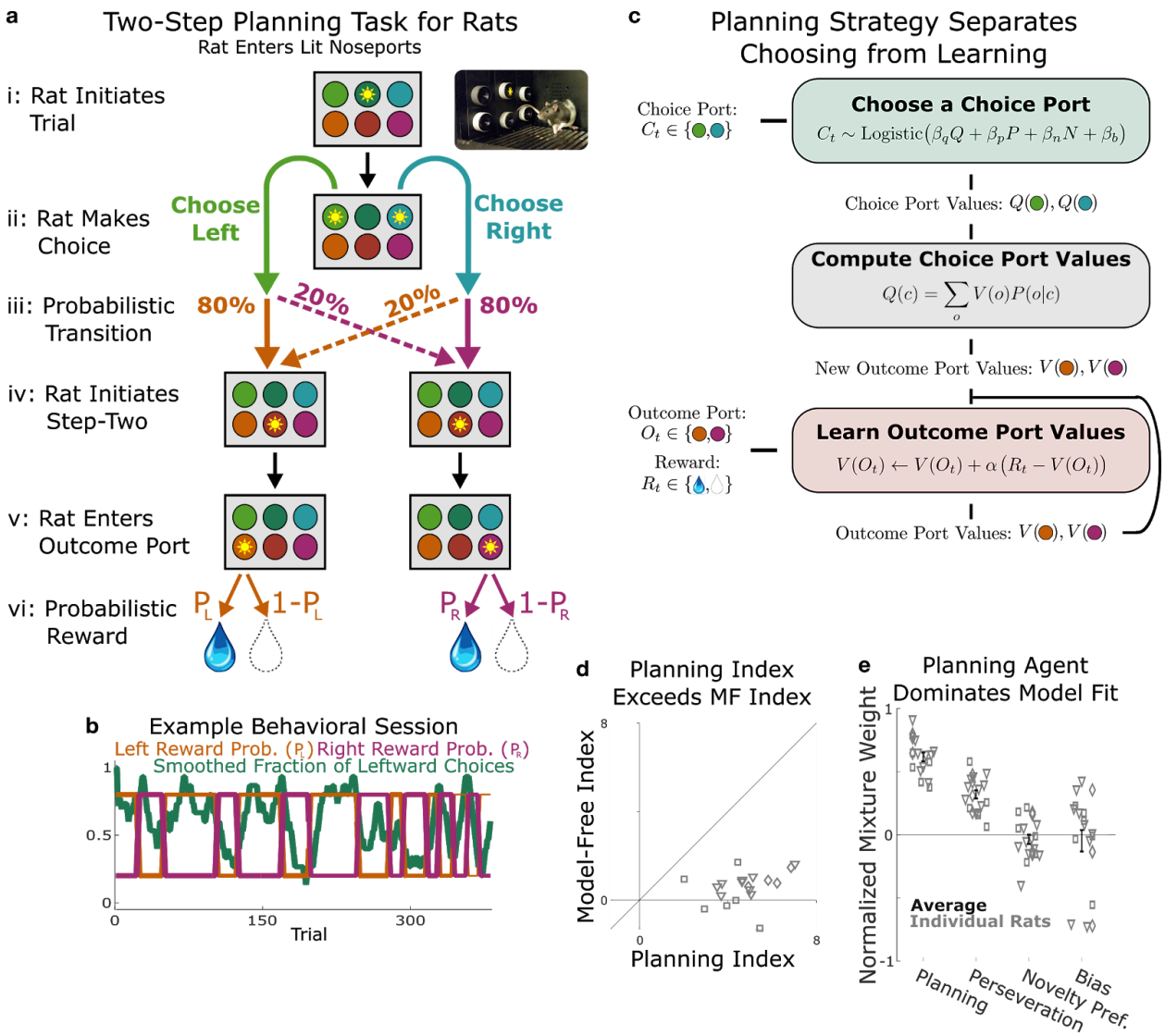

Figure 1. Two-step task separates choice values from outcome values. **a:** Rat two-step task. Rat initiates trial by entering the top center port (i), then chooses to enter one of two choice ports (ii). This leads to a probabilistic transition (iii) to one of two possible paths. In both paths, rat enters the bottom center port (v), causing one of two outcome ports to illuminate. Rat enters that outcome port (v), and receives a reward (vi). **b:** Example behavioral session. At unpredictable intervals, outcome port reward probabilities flip synchronously between high (80%) and low (20%). Rat adjusts choices accordingly **c:** Schematic of the planning strategy. Agent maintains value estimates (*V*) for each outcome port, based on a history of recent rewards at that port; as well as value estimates (*Q*) for each choice port, which are computed on each trial based on the outcome values and the world model (*P(o|c)*). Choices are drawn probabilistically, based on a weighted combination of these values and of the influence of three other behavioral patterns: perseveration, novelty preference, and bias (see Methods for details) **d:** Planning index and model-free index, shown for electrophysiology rats (n=6, squares), optogenetics rats (n=9, triangles), and sham optogenetics rats (n=4, diamonds). These indices are calculated using a trial-history regression analysis [see Methods; ref. 43] **e:** Weights of an agent-based model fit to rats' behavioral data [see Methods; ref. 43].

Neural Activity in OFC Codes Expected Value of Outcomes

Electrophysiological recordings during 51 behavioral sessions in six rats yielded 477 activity traces suitable for analysis, including both single- and multi-unit recordings (all results were similar across the two; Extended Data **Figs. S3**, **S4**). We sought to ask whether these units encoded three different types of value information, which we estimated trial-by-trial using fits of the agent-based model[45]. The first type was the difference in expected value between the left and the right choice ports, which we term "choice value difference". Representations of this kind would be consistent with a role in driving choice. The second was the expected value of the choice port that the rat actually selected on that trial, termed "chosen value". This signal has also been proposed to play a role in the choice process[49]. The last was the expected value of the outcome port visited by the rat on that trial, which we term "outcome value". Representations of outcome value are predictions about immediate reward probability, and are consistent with a role in learning.

Determining whether a unit codes these quantities requires separating their influence from that of other variables with which they may be correlated. For example, if the firing rate of a unit differs between left-choice and right-choice trials, it will also differ between trials in which the left port had a higher value and those in which the right port had a higher value (**Fig. 2a**). This happens because the rat is more likely to select the choice port that is currently higher-valued. Similarly, if the firing rate of a unit differs between rewarded and unrewarded trials, it will also differ between trials with a high and a low outcome port value (**Fig. 2c**). To quantify coding strength in light of these correlations (and others, **Fig. S5**), we used a multiple regression approach. Specifically, we fit a separate regression model to predict the spike counts of each unit in each of several time bins taken relative to the port entry events, using as regressors the potentially learning-related events from one trial (choice port, outcome port, reward, and their interactions, as well as outcome value) as well as the potentially choosing-related variables from the next trial (choice port, choice value difference, and chosen value). We then computed for each regressor the coefficient of partial determination (CPD; also known as "partial r-squared")[50,51], which quantifies the percentage of remaining variance explained by that regressor, once the influence of all other regressors has been accounted for (**Fig. 2**, bottom row).

Coefficients of partial determination can be computed for a particular fit (one unit in one time bin), or for a collection of fits (aggregating variance over units, bins, or both). First, we considered coding at the population level, computing CPD over all units for each time bin. We found robust coding of outcome-value, beginning at the time of entry into the bottom-center port, and peaking shortly after entry into the outcome port (**Fig 3a**) at 1.1%. In contrast, population coding of choice-value-difference and chosen-value was low in all time bins, reaching a maximum of only 0.52%. Population coding in the OFC was also present for other regressors in our model, especially for reward, choice, and outcome port (**Fig 3c**). Next, we considered coding in individual units, aggregating data from time bins within a one-second window around the time of each nose port entry. We found that a large fraction of units significantly modulated their firing rate according to outcome-value, with the largest fraction doing so at outcome port entry (175/477, 37%; permutation test at $p<0.01$). In contrast, a relatively small fraction of units modulated their firing rate according to choice-value-difference (largest at outcome port entry:

29/477, 6%), or to chosen-value (largest at choice port entry: 65/477, 14%). Furthermore, the magnitude of CPD in individual units was larger for outcome-value than for the other value regressors; considering the port entry event with the strongest coding for each regressor, the mean unit had CPD for outcome-value 3.8x larger than for choice-value-difference ($p=10^{-24}$, sign rank test; median unit 1.7x larger; **Fig. 3b**, note logarithmic axes), and 4.0x larger than for chosen-value ($p=10^{-14}$; median unit 1.5x; **Fig. 3b**; see also **Fig. S6**). Together, our results indicate that neural activity in OFC encodes the expected value of the visited outcome port much more strongly than it encodes either type of value information about the choice ports. In computational models (**Fig 1c**), this type of value information plays a role in supporting learning, but does not play a direct role in choice. Our neural recording results therefore suggest that the OFC may play a similar role in supporting learning, rather than a role in driving choice directly.

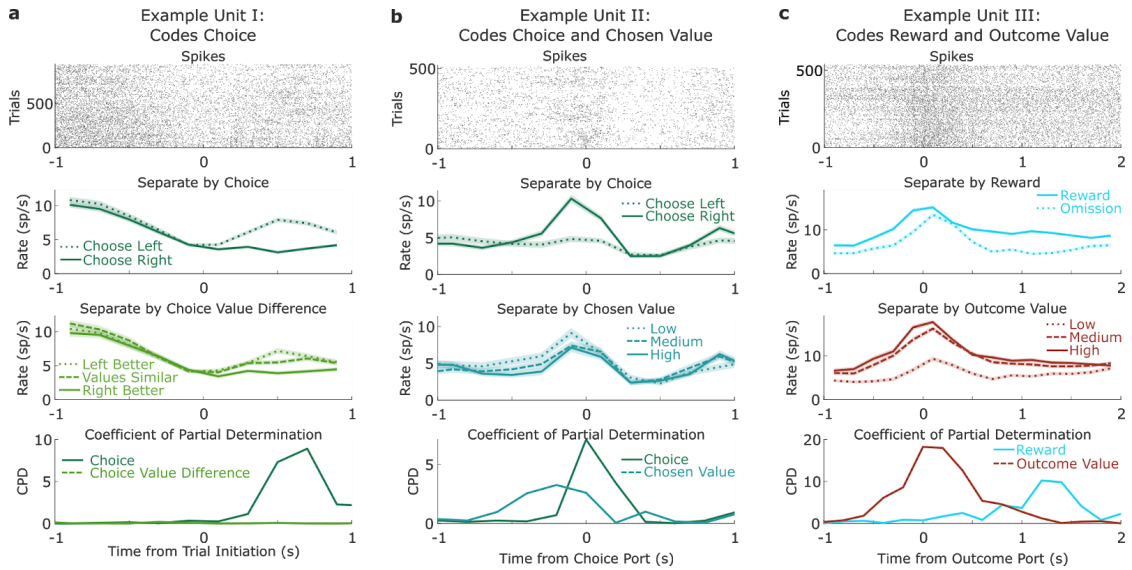

Figure 2. OFC units encode multiple correlated variables. a. Example unit whose firing rate differs both with the rat's choice and with the difference in value between the two possible choices. An analysis using coefficient of partial determination (CPD) reveals choice coding, but no coding of expected value. **b.** Example unit whose firing rate differs both with the rat's choice and with the expected value of the choice port visited on that trial. CPD analysis reveals coding of both of these variables, though with different timecourses. **c.** Example unit whose firing rate differs both with reward received and with the expected value of the outcome port visited. CPD analysis reveals coding of both of these variables, with different timecourses.

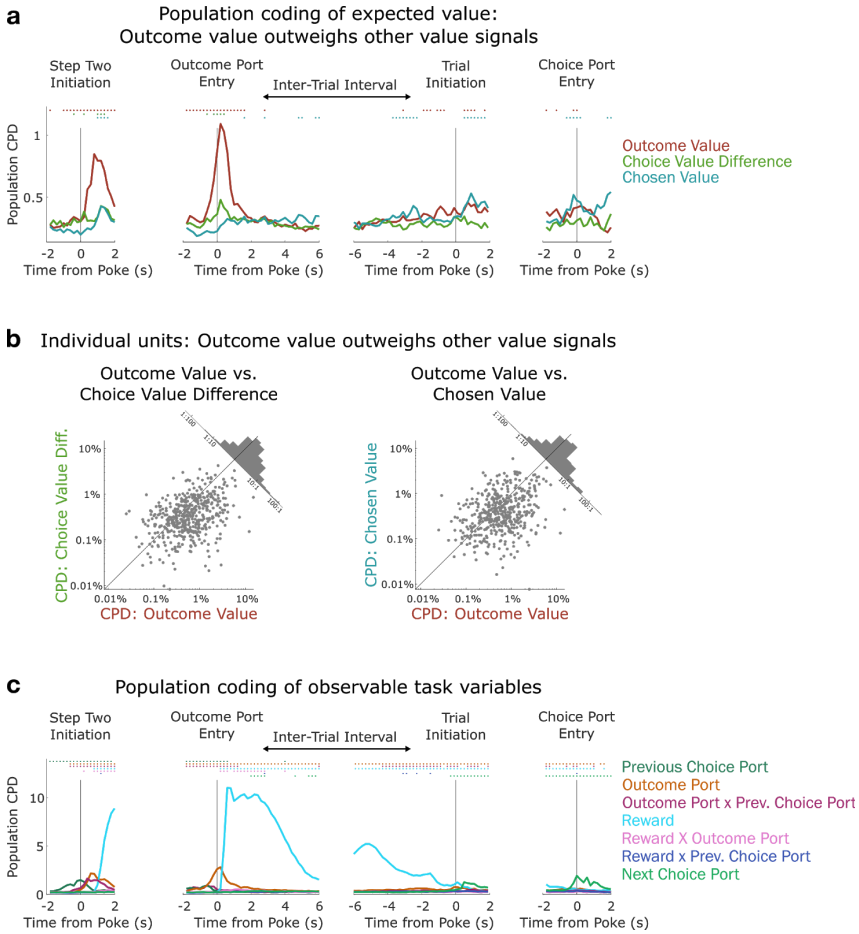

Figure 3. Coding of expected value of outcomes outweighs coding of expected value of choices. **a**. Timecourse of population CPD for the three expected value regressors. **b**. Left: Scatterplot showing CPD for each unit (n=477) for the outcome-value regressor against CPD for the choice-value-difference regressor, both computed in a one-second window centered on entry into the outcome port. Right: Scatterplot showing CPD for outcome-value, computed at outcome port entry, against CPD for the chosen-value regressor, computed at choice port entry. **c.** Timecourse of population CPD for the remaining regressors in the model, which reflect observable variables and interactions between them.

Inactivations of OFC Impair Update Process, not Choice Process

To assess the causal role of the OFC's value signals, we silenced neural activity using the optogenetic construct halorhodopsin (eNpHR3.0) during either the *outcome period* (beginning at entry into the outcome port and lasting until the end of reward consumption), the *choice period* (beginning at the end of reward consumption and lasting until entry into the choice port on the subsequent trial), or *both periods* (**Fig. 4a**). Previous work[43] had shown that whole-session pharmacological silencing of the OFC specifically attenuates the "planning index" (**Fig. 1e**, see Methods), which quantifies the extent to which the rats' choices are modulated by past trials' outcomes in a way consistent with planning. Here, we found that optogenetic inactivation spanning both the outcome and the reward period reproduced this effect, decreasing the planning index on the subsequent trial (p=0.007, t-test; **Fig 4b**). Inactivation during the outcome period alone also decreased the planning index on the subsequent trial (p=0.0007, **Fig 4b**), but

inactivation during the choice period alone did not (p=0.64). Comparing the strength of the effect across time periods, we found that both reward-period and both-periods inactivation produced effects that were similar to one another (p=0.5), but greater than the effect of choice-period inactivation (p=0.007, p=0.02). We repeated the experiment in a control group of four rats that were implanted with optical fibers but did not express halorhodopsin. This sham inactivation produced no significant effect on the planning index for any time period (all p>0.15; **Fig 4b**, grey diamonds), and experimental and control rats differed in the effects of outcome-period and both-period inactivation (p=0.02, p=0.02, two-sample t-tests). Together, these results confirm that silencing the OFC at the time of the outcome (and therefore the time of outcome-value coding, **Fig 2a**) is sufficient to disrupt planning behavior.

To assess which aspect of the behavior was affected by silencing the OFC, we perturbed our artificial planning agent in three separate ways. In the first, we disrupted choice-value representations (**Fig. 4d**, left; see Methods) on a subset of trials. In the second, we disrupted the representation of rewards actually delivered (**Fig. 4d**, middle). And in the third, we disrupted outcome-value representations (**Fig 4d**, right). Each of these produced a distinct pattern of behavior, most clearly visible when we separately computed the contribution to the planning index of the past three trials' rewards (**Fig. 4d**). Specifically, we found that scaling choice values on a particular trial affected the influence on choice of all past rewards (r_t, r_{t-1}, and r_{t-2}; **Fig 4d**, left), that scaling a particular trial's reward representation (r_t) affected only the influence of that reward (**Fig 4d**, middle), and that scaling outcome values on a particular trial (V) did not affect the influence of that trial's reward (r_t), but did affect the influence of rewards from previous trials (r_{t-1}, r_{t-2}; **Fig 4d** right). This is because this quantity acts as a summarized memory of the rewards of previous trials.

We performed the same analysis on the experimental data **(Fig. 4e**), and found that silencing OFC during the outcome period on a particular trial did not affect the influence of that trial's reward on the upcoming choice (p=0.2, paired t-test), but that it did affect the influence of the previous two trials' outcomes (p=0.004, p=0.02). This pattern was consistent with the synthetic dataset in which outcome value representations had been attenuated, but not the other synthetic datasets (compare **Fig. 4e** to **Fig. 4d**). We conclude that silencing the OFC in our task predominantly impairs outcome port value information, needed for updating value expectations, but has no effect on choice port value information, needed for driving choice.

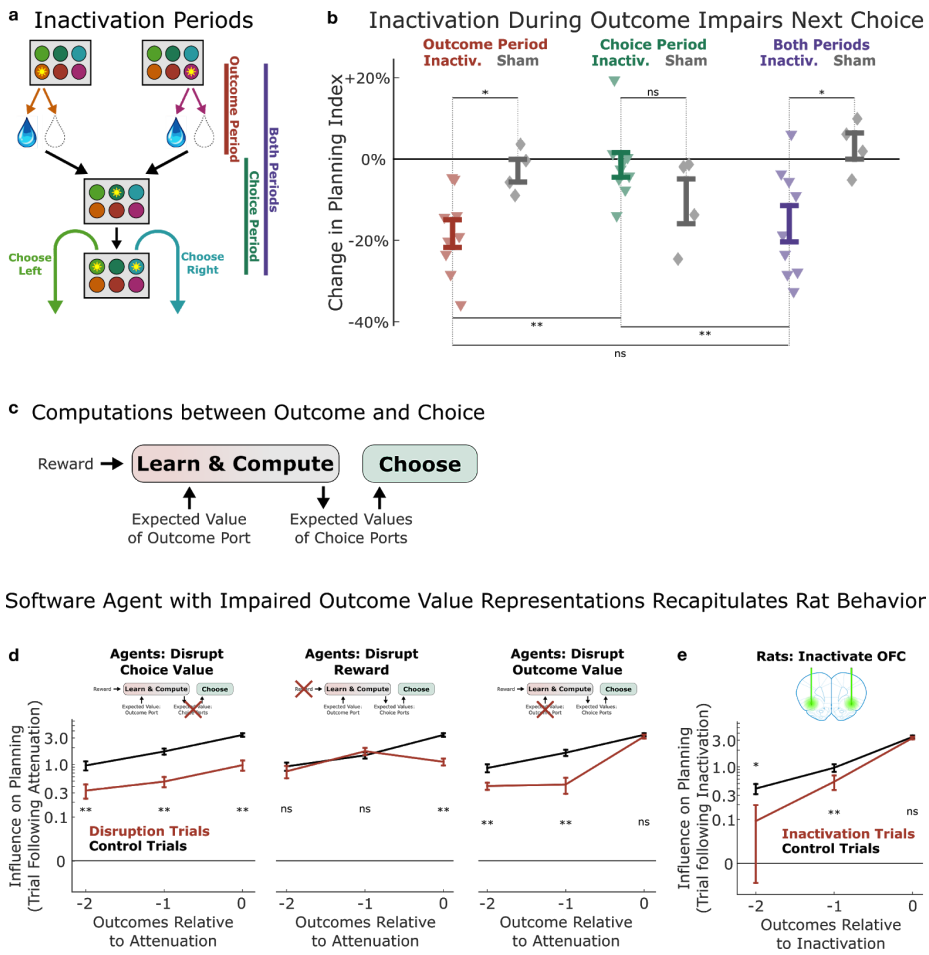

Figure 4. Inactivation of OFC attenuates influence of outcome values. **a:** Three time periods of inactivation. Outcome-period inactivation began when the rat entered the outcome port, and continued until the rat exited the port, or for a minimum of two seconds. Choice-period inactivation began after this outcome period, and continued until the rat entered the choice port on the next trial, or for a maximum of 15 seconds. Both-period inactivation encompassed both of these periods. **b:** Effects of inactivation on the planning index on the subsequent trial for experimental rats (n=9, colored triangles) and sham-inactivation rats (n=4, grey diamonds). Bars indicate standard errors across rats. **c:** Simplified schematic of the representations and computations that take place in our software agent between the delivery of the outcome on one trial and the choice on the next. Compare to Fig 1c. **d:** Analysis of synthetic datasets created by disrupting different representations within the software agent on a subset of trials. Each panel shows the contribution to the planning index of trial outcomes at different lags on choices, both on control trials (black) and on trials following disruption of a representation (red). Bars indicate standard error across simulated agents (see Methods) **e**: Same analysis as in c, applied to data from optogenetic inactivation of the OFC during the outcome period. Coronal section modified from [52]

Discussion

In formal models of value-based cognition, value representations can play two very different computational roles. First, they can drive choosing, as expected values of different available options are compared to one another, and the best is selected. Secondly, they can drive learning, as the expected value of an outcome is compared to the reward actually received, and future expectations are updated. Value representations in the OFC have been reported in many tasks and species, but it is still debated whether they drive one process, the other process, or both. The

rat two-step task gave us the opportunity to separate the two roles, both in terms of coding in neural activity and in terms of the behavioral impact of silencing OFC activity. We find limited representation of values associated with the available choices ("choice values"), little effect of silencing OFC at the putative time of choice, and effects of silencing inconsistent with impairing choice values in a computational model. Instead, we find strong representation of values associated with immediately impending reward outcomes ("outcome values"), a strong behavioral effect of silencing OFC at the time of those representations, and effects of silencing that are consistent with specifically impairing the use of outcome values for learning. These results suggest that value information in the rodent OFC does not drive choice directly, but instead provides information about expected outcomes to a separate learning process. In the computational model most consistent with our data, OFC inactivation does not prevent learning, but instead impairs one of the critical inputs to the learning process, namely expected outcome value. Our results therefore are consistent with the idea that learning does not take place in the OFC itself, but in some other structure to which OFC supplies this key input.

These results clarify the role of a neural signal that has been reported over a wide variety of tasks and species[4–7]. In decision tasks, several distinct types of value representation have been reported, but the most commonly observed is "chosen value"[27,29,51]: the expected reward associated with the option that the subject chose. The role of this signal is not clear, it has been interpreted both as representing a post-decision signal that provides input to a learning process[53], but also as an internal component of the choice process itself[49]. These roles are difficult to fully disentangle using classic tasks, in which both learning and choosing operate over the same set of items. Rat behavior in the two-step task separates these possibilities. If "chosen value" relates to the choice process, we would expect to see representations of the expected value of the port that was chosen (not learned about), while if it relates to the learning process, we would expect to see expected value of the port that was learned about (but not chosen). Our data predominantly provide evidence for the latter alternative in rats.

Our results build on prior work attempting to separate learning and choosing in trial-by-trial tasks in primates. One set of studies used a probabilistic reward-learning task (the "three-armed bandit" task), along with detailed trial-by-trial analysis to quantify deficits in "credit assignment", attributable to an impaired learning process, as well as deficits in decision-making[54]. Studies using this tool reported that both macaques and humans with lesions to the OFC showed impairments in credit assignment, while those with lesions to a neighboring region, ventromedial prefrontal cortex (vmPFC), showed impairments in decision-making[13,54,55]. A methodological issue with these studies is that they involved nonselective lesions, which damage both local neurons (from which electrophysiological signals are recorded) as well as fiber tracts that are merely passing through. A subsequent study used excitotoxic lesions, which do not damage passing fibers[15], and found that lesions spanning both OFC and vmPFC did not affect either credit assignment or decision making on this task. Instead, the study found that lesions of neighboring ventrolateral prefrontal cortex (vlPFC) caused impairments in credit assignment[15]. Together, these primate results suggest that neural activity in OFC itself is not necessary either for normal choosing or for normal learning on the three-armed bandit task, but that vlPFC may play a specific role in supporting learning[56]. Mapping these findings to the rodent brain is not entirely straightforward. While the distinction in rodents between OFC and

vmPFC is relatively clear, the distinction between OFC and vlPFC is much less so[57]. Anatomically, many of the criteria that motivate the proposed homology between rodent OFC (typically taken to mean areas LO and AIv) and primate OFC (typically taken to mean areas 11l and 13) may apply to parts of primate vlPFC (especially area 12) as well[57,58]. Like in OFC, neurons in vlPFC have been shown to represent expected outcomes[59,60], though these signals have not been characterized as intensively as those in primate OFC. A possibility is that primate OFC and vlPFC are specialized for functions that in rodents are both performed by the OFC.

Our results also build on those of previous studies that attempt to separate learning and choosing by separating these functions in time, typically by tens of minutes or longer. Here, however, we investigated a very different timescale in that learning and choosing were interleaved, both occurring on each trial of the two-step task. Our inactivation results are consistent with previous studies finding a role for the OFC in learning[17,19], but in apparent tension with others finding that inactivation impairs performance on a post-learning assay[14,19,21,38]. One possibility is that the OFC plays different roles for different timescales, or for these different types of behavior. Another possibility reconciling our results with previous studies is that the OFC plays a common role across timescales and tasks, supporting a process that updates choice mechanisms – in previous tasks this process may take place during either a learning or a probe session, while in our task it takes place between the outcome of one trial and the choice on the next.

Our results are in tension with those of some studies using economic choice tasks, in which subjects make decisions between pairs of well-learned stimuli leading to different quantities of differently flavored rewards. A pair of studies, one in mice[23] and one in primates[61], uses methods that disrupt OFC neural activity on a particular subset of trials, and reports that subjects' decisions are disrupted specifically on those trials. Recording studies in these tasks, both in primates and in mice[23,26], have reported neural correlates of the values of the individual options that are available. This "offer value" correlate is distinct from the "chosen value" correlate discussed earlier, and is suitable for a role in driving decisions directly. These results have led to the view that OFC value representations, at least in these tasks and species, plays a key role in decision-making[9,49]. Our results indicate that in the two-step task, OFC does not play this role, since "offer value" correlates are not present, and inactivation effects are inconsistent with a direct role in choice. One possible resolution to this tension is a difference between species (either biological or in terms of experimental implementations). In rats, silencing OFC does not cause impairments on an economic choice task[21,22]. Another possible resolution is a difference between tasks. It has been proposed that OFC plays a role specifically in choices between outcomes that offer different "goods" (e.g. different flavors)[23,62]. In primate tasks involving decisions between outcomes that differ in other properties – such as magnitude, probability, or delay – OFC value representations are typically more consistent with chosen-value than with offer-value coding[29,63,64]. Consistent with either of these possibilities are recent results from a choice task in rats in which options differed in probability and magnitude (but not flavor): rat OFC in this task did not carry correlates of offer value, and silencing impaired trial-history effects rather than the influence of offers on choice[35]. Another possibility reconciling our results with those from economic choice tasks that support a role for OFC in choosing is the idea that OFC contributes to a "policy update" process which can happen at different times in different tasks[39,40,65]. In our task, this process happens between the end of one trial and the beginning of the

next, and is interpreted as learning. In economic choice tasks, the update might happen immediately after the options are presented, but before a choice is made; in this scenario, it might be interpreted as related to choosing. Indeed, experiments analyzing response times[66], as well as moment-by-moment patterns of activity in the OFC[64], suggest that such a deliberative update process sometimes occurs, and that activity in the OFC may support it. This idea is broadly consistent with the proposal that OFC is critical for representing and using knowledge of latent task states, that is disambiguating states that are not perceptually distinct[67].

In summary, we find that rat behavior on a multi-step decision task dissociates the computational role of expected value information in choosing from its role in learning. We investigate the role of expected value representations in the OFC, and find a pattern of results indicating that these representations do not directly drive choice in this task, but instead support a learning process that updates choice mechanisms elsewhere in the brain. In addition to clarifying the function of the rodent OFC, these results provide insight into the overall architecture of value-based cognition. Specifically, they reveal that it is modular: that value-based learning and value-based choosing depend on separate neural representations. This lends credence to computational models which separate these functions[12,68]. It also motivates a search for the neural modules in the rat which play the remaining computational roles, including the role of directly driving choice.

Author Contributions

KJM, MMB, and CDB conceived the project. KJM designed the experiments and analysis, with supervision from MMB and CDB. KJM carried out all experiments and analyzed the data. KJM and CDB wrote the manuscript, based on a first draft by KJM, with extensive comments from MMB.

Acknowledgements

We thank Athena Akrami for assistance with array implant surgeries; Jovanna Teran, Klaus Osorio, Adrian Sirko, Samantha Stein, and Lillianne Teachen for assistance with animal training; and Peter Bibawi for help with histology. We thank Luca Mazzucato for identifying an error in an early version of our data processing pipeline. We thank Nathaniel Daw, Yael Niv, Geoffrey Schoenbaum, and Thomas Akam for helpful discussions, and Athena Akrami, Christine Constantinople, Nathaniel Daw, Cristina Domnisoru, Jeff Gauthier, Chuck Kopec, Olga Lositsky, Marcelo Mattar, Bas van Opheusden, Angela Radulescu, Ben Scott, Kim Stachenfeld, and Bob Wilson for helpful comments on the manuscript.

Methods

Subjects

All subjects were adult male Long-Evans rats (Taconic Biosciences; Hilltop Lab Animals), placed on a restricted water schedule to motivate them to work for water rewards. Rats were housed on a reverse 12-hour light cycle and trained during the dark phase of the cycle. Rats were pair housed during behavioral training and then single housed after being implanted with microwire arrays or optical fiber implants. Animal use procedures were approved by the Princeton University Institutional Animal Care and Use Committee and carried out in accordance with NIH standards.

Two-Step Behavioral Task

Rats were trained on a two-step behavioral task, following a shaping procedure which has been previously described[43]. Rats performed the task in custom behavioral chambers containing six "nose ports" arranged in two rows of three, each outfitted with a white LED for delivering visual stimuli, as well as an infrared LED and phototransistor for detecting rats' nose entries into the port. The left and right ports in the bottom row also contained sipper tubes for delivering water rewards. The rat initiated each trial by entering the illuminated top center port, causing the two top side ports ("choice ports") to illuminate. The rat then made his choice by entering one of these ports. Immediately upon entry into a choice port, two things happened: the bottom center port light illuminateed, and one of two possible sounds began to play, indicating which of the two bottom side ports ("outcome ports") would eventually be illuminated. The rat then entered the bottom center port, which caused the appropriate outcome port to illuminate. Finally, the rat entered the outcome port which illuminated, and received either a water reward or an omission. Once the rat had consumed the reward, a trial-end sound played, and the top center port illuminated again to indicate that the next trial was ready.

The selection of each choice port led to one of the outcome ports becoming available with 80% probability (common transition), and to the other becoming available with 20% probability (uncommon transition). These probabilities were counterbalanced across rats, but kept fixed for each rat for the entirety of his experience with the task. The probability that entry into each bottom side port would result in reward switched in blocks. In each block one port resulted in reward 80% of the time, and the other port resulted in reward 20% of the time. Block shifts happened unpredictably, with a minimum block length of 10 trials and a 2% probability of block change on each subsequent trial.

Analysis of Behavioral Data: Planning Index & Model-Free Index

We quantify the effect of past trials and their outcomes on future decisions using a logistic regression analysis based on previous trials and their outcomes[43,69]. We define vectors for each of the four possible trial outcomes: common-reward (CR), common-omission (CO), uncommon-reward (UR), and uncommon-omission (UO), each taking on a value of +1 for trials of their type where the rat selected the left choice port, a value of -1 for trials of their type where the rat selected the right choice port, and a value of 0 for trials of other types. We use the following regression model:

$$log\left(\frac{P_{left}(t)}{P_{right}(t)}\right) = \sum_{\tau=1}^{T}\beta_{CR}(\tau)\cdot CR(t-\tau) + \sum_{\tau=1}^{T}\beta_{CO}(\tau)\cdot CO(t-\tau) + \sum_{\tau=1}^{T}\beta_{UR}(\tau)\cdot UR(t-\tau) + \sum_{\tau=1}^{T}\beta_{UO}(\tau)\cdot UO(t-\tau) \quad (1)$$

where β_{cr}, β_{co}, β_{ur}, and β_{uo} are vectors of regression weights which quantify the tendency to repeat on the next trial a choice that was made τ trials ago and resulted in the outcome of their type, and T is a hyperparameter governing the number of past trials used by the model to predict upcoming choice, which was set to 3 for all analyses.

We expect model-free agents to repeat choices which lead to reward and switch away from those which lead to omissions[44], so we define a model-free index for a dataset as the sum of the appropriate weights from a regression model fit to that dataset:

$$ModelFreeIndex = \sum_{\tau=1}^{T} [\beta_{CR}(\tau) + \beta_{UR}(\tau)] - \sum_{\tau=1}^{T} [\beta_{UO}(\tau) + \beta_{CO}(\tau)] \quad (2)$$

We expect that planning agents will show the opposite pattern after uncommon transition trials, since the uncommon transition from one choice is the common transition from the other choice. We define a planning index:

$$PlanningIndex = \sum_{\tau=1}^{T} [\beta_{CR}(\tau) - \beta_{UR}(\tau)] + \sum_{\tau=1}^{T} [\beta_{UO}(\tau) - \beta_{CO}(\tau)] \quad (3)$$

We also compute the main effect of past choices on future choice:

$$PastChoices = \sum_{\tau=1}^{T} [\beta_{CR}(\tau) + \beta_{UR}(\tau) + \beta_{UO}(\tau) + \beta_{CO}(\tau)] \quad (4)$$

Mixture-of-Agents Behavior Model

We model behavior and obtain trial-by-trial estimates of value signals using an agent-based computational model which we have previously shown to provide a good explanation of rat behavior on the two-step task[43]. This model adopts the mixture-of-agents approach, in which each rat's behavior is described as resulting from the influence of a weighted average of several different "agents" implementing different behavioral strategies to solve the task. On each trial, each agent A computes a value, $Q_A(c)$, for each of the two available choices c, and the combined model makes a decision according to a weighted average of the various strategies' values, $Q_{total}(c)$:

$$Q_{total}(c) = \sum_{A \in \{agents\}} \beta_A Q_A(c)$$
$$\pi(c) = \frac{e^{Q_{total}(c)}}{\sum_{c'} e^{Q_{total}(c')}} \quad (5)$$

where the β's are weighting parameters determining the influence of each agent, and $\pi(c)$ is the probability that the mixture-of-agents will select choice c on that trial. The model which we have previously shown to provide the best explanation of rat's behavior contains four such agents: model-based temporal difference learning, novelty preference, perseveration, and bias.

Model-Based Temporal Difference Learning. Model-based temporal difference learning is a planning strategy, which maintains separate estimates of the probability with which each action (selecting the left or the right choice port) will lead to each outcome (the left or the right outcome port becoming available), $P(o|a)$, as well as the probability, $R(o)$, with which each outcome will lead to reward. This strategy assigns values to the actions by combining these probabilities to compute the expected probability with which selection of each action will ultimately lead to reward:

$$Q_{plan}(c) = \sum_{o} V(o) P(o|c) \quad (6)$$

At the beginning of each session, the reward estimate $V(o)$ is initialized to 0.5 for both outcomes, and the transition estimate $P(o|c)$ is set to the true transition function for the rat being modeled (0.8 for common and 0.2 for uncommon transitions). After each trial, the reward estimate for both outcomes is updated according to

$$V(o) \leftarrow \begin{cases} (1-\alpha) \cdot V(o) + \alpha \cdot r_t, \text{for } o = o_t \\ (1-\alpha) \cdot V(o) - \alpha \cdot r_t, \text{for } o \neq o_t \end{cases} \quad (7)$$

where o_t is the outcome that was observed on that trial, r_t is a binary variable indicating reward delivery, and α is a learning rate parameter constrained to lie between zero and one.

Novelty Preference. The novelty preference agent follows an "uncommon-stay/common switch" pattern, which tends to repeat choices when they lead to uncommon transitions on the previous trial, and to switch away from them when they lead to common transitions. Note that some rats have positive values of the β_{np} parameter weighting this agent (novelty preferring) while others have negative values (novelty averse; see **Fig 1e**):

$$Q_{np}(c_t) \leftarrow \begin{cases} 0, & \text{common transition trials} \\ 1, & \text{uncommon transition trials} \end{cases}$$
$$Q_{np}(c \neq c_t) \leftarrow 1 - Q_{np}(c_t) \tag{8}$$

Perseveration. Perseveration is a pattern which tends to repeat the choice that was made on the previous trial, regardless of whether it led to a common or an uncommon transition, and regardless of whether or not it led to reward.

$$Q_{persev}(c_t) \leftarrow 1$$
$$Q_{persev}(c \neq c_t) \leftarrow 0 \tag{9}$$

Bias. Bias is a pattern which tends to select the same choice port on every trial. Its value function is therefore static, with the extent and direction of the bias being governed by the magnitude and sign of this strategy's weighting parameter β_{bias}.

$$Q_{bias}(left) = 1$$
$$Q_{bias}(right) = -1 \tag{10}$$

Model Fitting

We implemented the model described above using the probabilistic programming language Stan[70,71], and performed maximum-a-posteriori fits using weakly informative priors on all parameters[72] The prior over the weighting parameters β was normal with mean 0 and sd 0.5, and the prior over α was a beta distribution with a=b=3. For ease of comparison, we normalize the weighting parameters β_{plan}, β_{np}, and β_{persev}, dividing each by the standard deviation of its agent's associated values (Q_{plan}, Q_{np}, and Q_{persev}) taken across trials. Since each weighting parameter affects behavior only by scaling the value output by its agent, this technique brings the weights into a common scale and facilitates interpretation of their relative magnitudes, analogous to the use of standardized coefficients in regression models.

Surgery: Microwire Array Implants

Six rats were implanted with microwire arrays (Tucker-David Technologies) targeting OFC unilaterally. Arrays contained tungsten microwires 4.5mm long and 50 µm in diameter, cut at a 60° angle at the tips. Wires were arranged in four rows of eight, with spacing 250 µm within-row and 375 µm between rows, for a total of 32 wires in a 1.125 mm by 1.75 mm rectangle. Target coordinates for the implant with respect to bregma were 3.1-4.2mm anterior, 2.4-4.2mm lateral, and 5.2mm ventral (~4.2mm ventral to brain surface at the posterior-middle of the array).

In order to expose enough of the skull for a craniotomy in this location, the jaw muscle was carefully resected from the lateral skull ridge in the area near the target coordinates. Dimpling of the brain surface was minimized following procedures described in more detail elsewhere[73]. Briefly, a bolus of petroleum jelly (Puralube, Dechra Veterinary Products) was placed in the center of the craniotomy to protect it, while cyanoacrylate glue (Vetbond, 3M) was used to adhere the pia mater to the skull at the periphery. The petroleum jelly was then removed, and the microwire array inserted slowly into the brain. Rats recovered for a minimum of one week, with ad lib access to food and water, before returning to training.

Electrophysiological Recordings

Once rats had recovered from surgery, recording sessions were performed in a behavioral chamber outfitted with a 32 channel recording system (Neuralynx). Spiking data was acquired using a bandpass filter between 600 and 6000 Hz and a spike detection threshold of 30 µV. Clusters were manually cut (Spikesort 3D, Neuralynx), and both single- and multi-units were considered.

Analysis of Electrophysiology Data

To determine the extent to which different variables were encoded in the neural signal, we fit a series of regression models to our spiking data. Models were fit to the spike counts emitted by each unit in 200 ms time bins taken relative to the four noseport entry events that made up each trial. There were ten total regressors, defined relative to a pair of adjacent trials. Seven of them were binary (coded as +-1), and related to observable task variables: the choice port selected on the earlier trial(left or right), the outcome port visited (left or right), the reward received (reward or omission), the interaction between choice port and outcome port (common or uncommon transition), the interaction between choice port and reward, the interaction between outcome port and reward, and the choice port selected on the later trial. Three additional regressors were continuous and related to subjective reward expectation: the expected value of the outcome port visited (*V*) for the first trial, the value difference between the choice ports (*Q(left)* - *Q(right)*), and the value of the choice port selected (*Q(chosen)*) on the subsequent trial. These last three regressors were obtained using the agent-based computational model described above, with parameters fit separately to each rat's behavioral data. Regressors were z-scored to facilitate comparison of fit regression weights. Models were fit using the Matlab function `glmnet`[74] using a Poisson noise model. We fit each model both with no regularization and with L1 regularization ($\alpha = 1$, $\lambda = 0, 10^{-10}, 10^{-9}, \ldots, 10^{-1}$), using the model with the weakest regularization that still allowed all weights to be identifiable.

In our task, many of these regressors were correlated with one another (**Fig. S3**), so we quantify encoding using the coefficient of partial determination (CPD; also known as partial r-squared) associated with each[50,51]. This measure quantifies the fraction of variance explained by each regressor, once the variance explained by all other regressors has been taken account of:

$$CPD(X_i, u, t) = \frac{\Big(SSE(X_{-i}, u, t) - SSE(X_{all}, u, t)\Big)}{SSE(X_{-i}, u, t)} \tag{11}$$

where *u* refers to a particular unit, *t* refers to a particular time bin, and SSE(X_{all}) refers to the sum-squared-error of a regression model considering all eight regressors described above, and SSE(X_{-i}) refers to the sum-squared-error of a model considering the seven regressors other than X_i. We compute total CPD for each unit by summing the SSE associated with the regression models for that unit for all time bins:

$$CPD(X_i, u) = \frac{\sum_t \Big(SSE(X_{-i}, u, t) - SSE(X_{all}, u, t)\Big)}{\sum_t SSE(X_{-i}, u, t)} \tag{12}$$

We report this measure both for individual units and bins (Fig 2, bottom), as well as for the case where the sum is taken over the five bins making up a 1s time window centered on a particular port entry event (top neutral center port, choice port, bottom neutral center port, or outcome port). Cells were labeled as significantly encoding a regressor if the CPD for that regressor exceeded that of more than 99% of CPDs computed based on datasets with circularly shifted trial labels. We use permuted, rather than shuffled, labels in order to preserve trial-by-trial correlational structure. Differences in the strength of encoding of different regressors were assessed using a signrank test on the differences in CPD.

We compute total population CPD for a particular time bin in an analogous way:

$$CPD(X_i, t) = \frac{\sum_u \Big(SSE(X_{-i}, u, t) - SSE(X_{all}, u, t)\Big)}{\sum_u SSE(X_{-i}, u, t)} \tag{13}$$

Time bins were labeled as significantly encoding a regressor if the population CPD for that time bin exceeded the largest population CPD for any regressor in more than 99% of the datasets with permuted trial labels. We compute total CPD for each regressor by summing SSE over both units and time bins, and assess the significance of differences in this quantity by comparing them to differences in total CPD computed for the shuffled datasets.

Surgery: Optical Fiber Implant and Virus Injection

Rats were implanted with sharpened fiber optics and received virus injections following procedures similar to those described previously[73,75,76], and documented in detail on the Brody lab website (http://brodywiki.princeton.edu/wiki/index.php/Etching_Fiber_Optics). A 50/125 µm LC-LC duplex fiber cable (Fiber Cables) was dissected to produce four blunt fiber segments with LC connectors. These segments were then sharpened by immersing them in hydroflouric acid and slowly retracting them using a custom-built motorized jig attached to a micromanipulator (Narashige International) holding the fiber. Each rat was implanted with two sharpened fibers, in order to target OFC bilaterally. Target coordinates with respect to bregma were 3.5mm anterior, 2.5mm lateral, 5mm ventral. Fibers were angled 10 degrees laterally, to make space for female-female LC connectors which were attached to each and included as part of the implant.

Four rats were implanted with sharpened optical fibers only, but received no injection of virus. These rats served as uninfected controls.

Nine additional rats received both fiber implants as well as injections of a virus (AAV5-CaMKIIα-eNpHR3.0-eYFP; UNC Vector Core) into the OFC to drive expression of the light-activated inhibitory opsin eNpHR3.0. Virus was loaded into a glass micropipette mounted into a Nanoject III (Drummond Scientific), which was used for injections. Injections involved five tracks arranged in a plus-shape, with spacing 500µm. The center track was located 3.5mm anterior and 2.5mm lateral to bregma, and all tracks extended from 4.3 to 5.7mm ventral to bregma. In each track, 15 injections of 23 nL were made at 100µm intervals, pausing for ten seconds between injections, and for one minute at the bottom of each track. In total 1.7 µl of virus were delivered to each hemisphere over a period of about 20 minutes.

Rats recovered for a minimum of one week, with ad lib access to food and water, before returning to training. Rats with virus injections returned to training, but did not begin inactivation experiments until a minimum of six weeks had passed, to allow for virus expression.

Optogenetic Perturbation Experiments

During inactivation experiments, rats performed the task in a behavioral chamber outfitted with a dual fiber optic patch cable connected to a splitter and a single-fiber commutator (Princetel) mounted in the ceiling. This fiber was coupled to a 200 mW 532 nm laser (OEM Laser Systems) under the control of a mechanical shutter (ThorLabs) by way of a fiber port (ThorLabs). The laser power was tuned such that each of the two fibers entering the implant received between 25 and 30 mW of light when the shutter was open.

Each rat received several sessions in which the shutter remained closed, in order to acclimate to performing the task while tethered. Once the rat showed behavioral performance while tethered that was similar to his performance before the implant surgery, inactivation sessions began. During these sessions, the laser shutter was opened (causing light to flow into the implant, activating the eNpHR3.0 and silence neural activity) on 7% of trials each in one of three time periods. “Outcome period” inactivation began when the rat entered the bottom center port at the end of the trial, and ended either when the rat had left the port and remained out for a minimum of 500 ms, or after 2.5 s. “Choice period” inactivation began at the end of the outcome period and lasted until the rat entered the choice port on the following trial. “Both period” inactivation encompassed both the outcome period and the choice period. The total duration of the inactivation therefore depended in part on the movement times of the rat, and was somewhat variable from trials to trial (**Fig ED8**). If a scheduled inactivation would last more than 15 s, inactivation was terminated, and that trial was excluded from analysis. Due to constraints of the bControl software, inactivation was only performed on even-numbered trials.

Analysis of Optogenetic Effects on Behavior

We quantify the effects of optogenetic inhibition on behavior by computing separately the planning index for trials following inactivation of each type (outcome period, choice period, both periods) and for control trials. Specifically, we fit the trial history regression model of Equation 1 with a separate set of weights for trials following inactivation of each type:

$$log\left(\frac{P_{left}(t)}{P_{right}(t)}\right) = \sum_{\tau=1}^{T}\beta_{CR,i}(\tau)\cdot CR(t-\tau) + \sum_{\tau=1}^{T}\beta_{CO,i}(\tau)\cdot CO(t-\tau) + \sum_{\tau=1}^{T}\beta_{UR,i}(\tau)\cdot UR(t-\tau) + \sum_{\tau=1}^{T}\beta_{UO,i}(\tau)\cdot UO(t-\tau) \tag{14}$$

$$i = \begin{cases} cntrl, & \text{trial } t-1 \text{ was control trial} \\ out, & \text{trial } t-1 \text{ had outcome period inactivation} \\ ch, & \text{trial } t-1 \text{ had choice period inactivation} \\ both, & \text{trial } t-1 \text{ had both} \end{cases} \tag{15}$$

We used maximum a posteriori fitting in which the priors were Normal(0,1) for weights corresponding to control trials, and Normal($\beta_{X,\ cntrl}$, 1) for weights corresponding to inactivation trials, where $\beta_{X,\ cntrl}$ is the corresponding control trial weight – e.g. the prior for $\beta_{CR,\ out}(1)$ is Normal($\beta_{CR,\ cntrl}(1)$, 1). This prior embodies the belief that the effect of inactivation on behavior is likely to be small, and that the direction of any effect is equally likely to be positive or negative. This ensures that our priors cannot induce any spurious differences between control and inactivation conditions into the parameter estimates. We then compute a planning index separately for the weights of each type, modifying equation 3:

$$PlanningIndex_i = \sum_{\tau=1}^{T}\left[\beta_{CR,i}(\tau) - \beta_{UR,i}(\tau)\right] + \sum_{\tau=1}^{T}\left[\beta_{UO,i}(\tau) - \beta_{CO,i}(\tau)\right] \tag{16}$$

We compute the relative change in planning index for each inactivation condition: ($PlanningIndex_i$ - $PlanningIndex_{cntrl}$) / $PlanningIndex_{cntrl}$, and report three types of significance tests on this quantity. First, we test for each inactivation condition the hypothesis that there was a significant change in the planning index, reporting the results of a one-sample t-test over rats. Next, we test the hypothesis that different inactivation conditions had effects of different sizes on the planning index, reporting a paired t-test over rats. Finally, we test the hypothesis for each condition that inactivation had a different effect than sham inactivation (conducted in rats which had not received virus injections to deliver eNpHR3.0), reporting a two-sample t-test.

To test the hypothesis that inactivation specifically impairs the effect of distant past outcomes on upcoming choice, we break down the planning index for each condition by the index of the weights the contribute to it:

$$PlanningIndex_i(\tau) = \left[\beta_{CR,i}(\tau) - \beta_{UR,i}(\tau)\right] + \left[\beta_{UO,i}(\tau) - \beta_{CO,i}(\tau)\right] \tag{17}$$

We report these trial-lagged planning indices for each inactivation condition, and assess the significance of the difference between inactivation and control conditions at each lag using a paired t-test.

Synthetic Datasets

To generate synthetic datasets for comparison to optogenetic inactivation data, we generalized the behavioral model to separate the contributions of representations of expected value and of immediate reward. In particular, we replaced the learning equation within the model-based RL agent (Equation 7) with the following:

$$V(o) \leftarrow \begin{cases} \alpha_{value}\cdot V(o) + \alpha_{reward}\cdot R_t + (1-\alpha_{value}-\alpha_{reward})\cdot \mathrm{E}[V], \text{for } o = o_t \\ \alpha_{value}\cdot V(o) - \alpha_{reward}\cdot R_t + (1-\alpha_{value}-\alpha_{reward})\cdot \mathrm{E}[V], \text{for } o \neq o_t \end{cases} \tag{18}$$

where α_{value} and α_{reward} are separate learning rate parameters, constrained to be nonnegative and to have a sum no larger than one, and E[V] represents the expected reward of a random-choice policy on the task, which in the case of our task is equal to 0.5.

To generate synthetic datasets in which silencing the OFC impairs choice value representations, outcome value representations, or reward representations, we decrease the parameter β_{plan}, α_{value}, or α_{reward}, respectively. Specifically, we first fit the model to the dataset for each rat in the optogenetics experiment (n=9) as above (i.e. using equation 7 as the learning rule) to obtain maximum a posteriori parameters. We translated these parameters to the optogenetics (equation 18) version of the model by setting α_{value} equal to the fit parameter α and α_{reward} equal to 1 - α. We then generated four synthetic datasets for each rat. For the control dataset, the fit parameters were used on trials of all types, regardless of whether inhibition of OFC was scheduled on that trial. For the "impaired outcome values" dataset, α_{value} was decreased specifically for trials with inhibition scheduled during the outcome period or both periods, but not on trials with inhibition during the choice period or on control trials. For the "impaired reward processing" dataset, α_{reward} was decreased on these trials instead. For the "impaired decision-making" dataset, β_{plan} was decreased specifically on trials following inhibition. In all cases, the parameter to be decreased was multiplied by 0.3, and synthetic datasets consisted of 100,000 total trials per rat.

Histological Verification of Targeting

We verified that surgical implants were successfully placed in the OFC using standard histological techniques. At experimental endpoint, rats with electrode arrays were anesthetized, and microlesions were made at the site of each electrode tip by passing current through the electrodes. Rats were then perfused transcardially with saline followed by formalin. Brains were sliced using a vibratome and imaged using an epifluorescent microscope. Recording sites were identified using these microlesions and the scars created by the electrodes in passing, as well dimples in the surface of the brain. Locations of optical fibers were identified using the scars created by their passage. Location of virus expression was identified by imaging the YFP conjugated to the eNpHR3.0 molecule. **Figure S7**

Data and Code Availability

All data collected for the purpose of this paper, and all software used in the analysis of this data, are available from the corresponding author upon request. Software used for training rats is available on the Brody lab website.

Citations

1. Sugrue, L. P., Corrado, G. S. & Newsome, W. T. Choosing the greater of two goods: neural currencies for valuation and decision making. *Nat. Rev. Neurosci.* **6**, 363–375 (2005).
2. Daw, N. D. & O'Doherty, J. P. Chapter 21 - Multiple Systems for Value Learning. in *Neuroeconomics (Second Edition)* (eds. Glimcher, P. W. & Fehr, E.) 393–410 (Academic Press, 2014).
3. Lee, D., Seo, H. & Jung, M. W. Neural basis of reinforcement learning and decision making. *Annu. Rev. Neurosci.* **35**, 287–308 (2012).
4. O'Doherty, J. P. Lights, camembert, action! The role of human orbitofrontal cortex in encoding stimuli, rewards, and choices. *Ann. N. Y. Acad. Sci.* **1121**, 254–272 (2007).
5. Rudebeck, P. H. & Murray, E. A. The orbitofrontal oracle: cortical mechanisms for the prediction and evaluation of specific behavioral outcomes. *Neuron* **84**, 1143–1156 (2014).
6. Wallis, J. D. Cross-species studies of orbitofrontal cortex and value-based decision-making. *Nat. Neurosci.* **15**, 13–19 (2012).
7. Stalnaker, T. A., Cooch, N. K. & Schoenbaum, G. What the orbitofrontal cortex does not do. *Nat. Neurosci.* **18**, 620–627 (2015).
8. Wallis, J. D. Orbitofrontal cortex and its contribution to decision-making. *Annu. Rev. Neurosci.* (2007).
9. Padoa-Schioppa, C. & Conen, K. E. Orbitofrontal Cortex: A Neural Circuit for Economic Decisions. *Neuron* **96**, 736–754 (2017).
10. Schoenbaum, G., Roesch, M. R., Stalnaker, T. A. & Takahashi, Y. K. A new perspective on the role of the orbitofrontal cortex in adaptive behaviour. *Nat. Rev. Neurosci.* **10**, 885–892 (2009).
11. Walton, M. E., Behrens, T. E. J., Noonan, M. P. & Rushworth, M. F. S. Giving credit where credit is due: orbitofrontal cortex and valuation in an uncertain world. *Ann. N. Y. Acad. Sci.* **1239**, 14–24

(2011).

12. Song, H. F., Yang, G. R. & Wang, X.-J. Reward-based training of recurrent neural networks for cognitive and value-based tasks. *Elife* **6**, (2017).

13. Noonan, M. P. *et al.* Separate value comparison and learning mechanisms in macaque medial and lateral orbitofrontal cortex. *Proc. Natl. Acad. Sci. U. S. A.* **107**, 20547–20552 (2010).

14. Murray, E. A., Moylan, E. J., Saleem, K. S., Basile, B. M. & Turchi, J. Specialized areas for value updating and goal selection in the primate orbitofrontal cortex. *Elife* **4**, (2015).

15. Rudebeck, P. H., Saunders, R. C., Lundgren, D. A. & Murray, E. A. Specialized Representations of Value in the Orbital and Ventrolateral Prefrontal Cortex: Desirability versus Availability of Outcomes. *Neuron* **95**, 1208–1220.e5 (2017).

16. Gallagher, M., McMahan, R. W. & Schoenbaum, G. Orbitofrontal cortex and representation of incentive value in associative learning. *J. Neurosci.* **19**, 6610–6614 (1999).

17. Takahashi, Y. K. *et al.* The orbitofrontal cortex and ventral tegmental area are necessary for learning from unexpected outcomes. *Neuron* **62**, 269–280 (2009).

18. McDannald, M. A., Lucantonio, F., Burke, K. A., Niv, Y. & Schoenbaum, G. Ventral striatum and orbitofrontal cortex are both required for model-based, but not model-free, reinforcement learning. *J. Neurosci.* **31**, 2700–2705 (2011).

19. Jones, J. L. *et al.* Orbitofrontal cortex supports behavior and learning using inferred but not cached values. *Science* **338**, 953–956 (2012).

20. Sadacca, B. F. *et al.* Orbitofrontal neurons signal sensory associations underlying model-based inference in a sensory preconditioning task. *Elife* **7**, (2018).

21. Gardner, M. P. H., Conroy, J. S., Shaham, M. H., Styer, C. V. & Schoenbaum, G. Lateral Orbitofrontal Inactivation Dissociates Devaluation-Sensitive Behavior and Economic Choice. *Neuron* **96**, 1192–1203.e4 (2017).

22. Gardner, M. P. H., Conroy, J. C., Sanchez, D. C., Zhou, J. & Schoenbaum, G. Real-Time Value Integration during Economic Choice Is Regulated by Orbitofrontal Cortex. *Curr. Biol.* **29**, 4315–4322.e4 (2019).

23. Kuwabara, M., Kang, N., Holy, T. E. & Padoa-Schioppa, C. Neural mechanisms of economic choices in mice. *Elife* **9**, (2020).

24. Schoenbaum, G., Chiba, A. A. & Gallagher, M. Orbitofrontal cortex and basolateral amygdala encode expected outcomes during learning. *Nat. Neurosci.* **1**, 155–159 (1998).

25. Gottfried, J. A., O'Doherty, J. & Dolan, R. J. Encoding predictive reward value in human amygdala and orbitofrontal cortex. *Science* **301**, 1104–1107 (2003).

26. Padoa-Schioppa, C. & Assad, J. A. Neurons in the orbitofrontal cortex encode economic value. *Nature* **441**, 223–226 (2006).

27. Sul, J. H., Kim, H., Huh, N., Lee, D. & Jung, M. W. Distinct roles of rodent orbitofrontal and medial prefrontal cortex in decision making. *Neuron* **66**, 449–460 (2010).

28. Thorpe, S. J., Rolls, E. T. & Maddison, S. The orbitofrontal cortex: neuronal activity in the behaving monkey. *Exp. Brain Res.* **49**, 93–115 (1983).

29. Wallis, J. D. & Miller, E. K. Neuronal activity in primate dorsolateral and orbital prefrontal cortex during performance of a reward preference task. *Eur. J. Neurosci.* **18**, 2069–2081 (2003).

30. Costa, V. D. & Averbeck, B. B. Primate orbitofrontal cortex codes information relevant for managing explore-exploit tradeoffs. *J. Neurosci.* (2020) doi:10.1523/JNEUROSCI.2355-19.2020.

31. Tremblay, L. & Schultz, W. Relative reward preference in primate orbitofrontal cortex. *Nature* **398**, 704–708 (1999).

32. Kennerley, S. W., Dahmubed, A. F., Lara, A. H. & Wallis, J. D. Neurons in the frontal lobe encode the value of multiple decision variables. *J. Cogn. Neurosci.* **21**, 1162–1178 (2009).

33. Rudebeck, P. H., Mitz, A. R., Chacko, R. V. & Murray, E. A. Effects of amygdala lesions on

reward-value coding in orbital and medial prefrontal cortex. *Neuron* **80**, 1519–1531 (2013).

34. Blanchard, T. C., Hayden, B. Y. & Bromberg-Martin, E. S. Orbitofrontal cortex uses distinct codes for different choice attributes in decisions motivated by curiosity. *Neuron* **85**, 602–614 (2015).
35. Constantinople, C. M. *et al.* Lateral orbitofrontal cortex promotes trial-by-trial learning of risky, but not spatial, biases. *Elife* **8**, (2019).
36. Padoa-Schioppa, C. Neuronal origins of choice variability in economic decisions. *Neuron* **80**, 1322–1336 (2013).
37. Lak, A. *et al.* Reinforcement biases subsequent perceptual decisions when confidence is low, a widespread behavioral phenomenon. *Elife* **9**, (2020).
38. Gremel, C. M. & Costa, R. M. Orbitofrontal and striatal circuits dynamically encode the shift between goal-directed and habitual actions. *Nat. Commun.* **4**, 2264 (2013).
39. Gershman, S. J., Markman, A. B. & Otto, A. R. Retrospective revaluation in sequential decision making: a tale of two systems. *J. Exp. Psychol. Gen.* **143**, 182–194 (2014).
40. Ludvig, E. A., Mirian, M. S., Kehoe, E. & Sutton, R. S. Associative Learning from Replayed Experience. *bioRxiv* 100800 (2017) doi:10.1101/100800.
41. Lombrozo, T. 'Learning by Thinking' in Science and in Everyday Life. in *The scientific imagination* (Oxford University Press, 2017).
42. McDaniel, M. A. & Einstein, G. O. *Prospective Memory: An Overview and Synthesis of an Emerging Field*. (SAGE Publications, 2007).
43. Miller, K. J., Botvinick, M. M. & Brody, C. D. Dorsal hippocampus contributes to model-based planning. *Nat. Neurosci.* **20**, 1269–1276 (2017).
44. Daw, N. D., Gershman, S. J., Seymour, B., Dayan, P. & Dolan, R. J. Model-based influences on humans' choices and striatal prediction errors. *Neuron* **69**, 1204–1215 (2011).
45. Daw, N. D. Trial-by-trial data analysis using computational models. in *Decision Making, Affect, and*

Learning 3–38 (2011).

46. Daw, N. D., Niv, Y. & Dayan, P. Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioral control. *Nat. Neurosci.* **8**, 1704–1711 (2005).
47. Dolan, R. J. & Dayan, P. Goals and habits in the brain. *Neuron* **80**, 312–325 (2013).
48. Sutton, R. S. & Barto, A. G. Reinforcement Learning: An Introduction (second edi ed.). (2017).
49. Rustichini, A. & Padoa-Schioppa, C. A neuro-computational model of economic decisions. *J. Neurophysiol.* **114**, 1382–1398 (2015).
50. Cai, X., Kim, S. & Lee, D. Heterogeneous coding of temporally discounted values in the dorsal and ventral striatum during intertemporal choice. *Neuron* **69**, 170–182 (2011).
51. Kennerley, S. W., Behrens, T. E. J. & Wallis, J. D. Double dissociation of value computations in orbitofrontal and anterior cingulate neurons. *Nat. Neurosci.* **14**, 1581–1589 (2011).
52. Paxinos, G. & Watson, C. *The Rat Brain in Stereotaxic Coordinates: Hard Cover Edition*. (Elsevier, 2006).
53. Schoenbaum, G., Takahashi, Y., Liu, T.-L. & McDannald, M. A. Does the orbitofrontal cortex signal value? *Ann. N. Y. Acad. Sci.* **1239**, 87–99 (2011).
54. Walton, M. E., Behrens, T. E. J., Buckley, M. J., Rudebeck, P. H. & Rushworth, M. F. S. Separable learning systems in the macaque brain and the role of orbitofrontal cortex in contingent learning. *Neuron* **65**, 927–939 (2010).
55. Noonan, M. P., Chau, B. K. H., Rushworth, M. F. S. & Fellows, L. K. Contrasting Effects of Medial and Lateral Orbitofrontal Cortex Lesions on Credit Assignment and Decision-Making in Humans. *J. Neurosci.* **37**, 7023–7035 (2017).
56. Murray, E. A. & Rudebeck, P. H. Specializations for reward-guided decision-making in the primate ventral prefrontal cortex. *Nat. Rev. Neurosci.* **19**, 404–417 (2018).
57. Price, J. L. Definition of the orbital cortex in relation to specific connections with limbic and visceral

structures and other cortical regions. *Ann. N. Y. Acad. Sci.* **1121**, 54–71 (2007).

58. Schoenbaum, G., Setlow, B. & Gallagher, M. Orbitofrontal cortex: modeling prefrontal function in rats. *The neuropsychology of memory* **3**, 463–477 (2002).

59. Rich, E. L. & Wallis, J. D. Medial-lateral organization of the orbitofrontal cortex. *J. Cogn. Neurosci.* **26**, 1347–1362 (2014).

60. Kobayashi, S., Pinto de Carvalho, O. & Schultz, W. Adaptation of reward sensitivity in orbitofrontal neurons. *J. Neurosci.* **30**, 534–544 (2010).

61. Ballesta, S., Shi, W., Conen, K. E. & Padoa-Schioppa, C. Values Encoded in Orbitofrontal Cortex Are Causally Related to Economic Choices. *bioRxiv* (2020).

62. Padoa-Schioppa, C. Neurobiology of economic choice: a good-based model. *Annu. Rev. Neurosci.* **34**, 333–359 (2011).

63. Roesch, M. R. & Olson, C. R. Neuronal activity related to reward value and motivation in primate frontal cortex. *Science* **304**, 307–310 (2004).

64. Rich, E. L. & Wallis, J. D. Decoding subjective decisions from orbitofrontal cortex. *Nat. Neurosci.* **19**, 973–980 (2016).

65. Mattar, M. G. & Daw, N. D. Prioritized memory access explains planning and hippocampal replay. *Nat. Neurosci.* **21**, 1609–1617 (2018).

66. Krajbich, I., Armel, C. & Rangel, A. Visual fixations and the computation and comparison of value in simple choice. *Nat. Neurosci.* **13**, 1292–1298 (2010).

67. Wilson, R. C., Takahashi, Y. K., Schoenbaum, G. & Niv, Y. Orbitofrontal cortex as a cognitive map of task space. *Neuron* **81**, 267–279 (2014).

68. Joel, D., Niv, Y. & Ruppin, E. Actor – critic models of the basal ganglia: new anatomical and computational perspectives. *Neural Netw.* **15**, 535–547 (2002).

69. Lau, B. & Glimcher, P. W. Dynamic response-by-response models of matching behavior in rhesus

monkeys. *J. Exp. Anal. Behav.* **84**, 555–579 (2005).

70. Stan Development Team. MatlabStan: The MATLAB interface to Stan. (2016).

71. Bob Carpenter, Andrew Gelman, Matt Hoffman, Daniel Lee, Ben Goodrich, Michael Betancourt, Michael A. Brubaker, Jiqiang Guo, Peter Li, and Allen Riddell. Stan: A Probabilistic Programming Language. (2016).

72. Gelman, A. *et al. Bayesian Data Analysis, Third Edition*. (CRC Press, 2013).

73. Akrami, A., Kopec, C. D., Diamond, M. E. & Brody, C. D. Posterior parietal cortex represents sensory stimulus history and is necessary for its effects on behavior. *bioRxiv* 182246 (2017) doi:10.1101/182246.

74. Qian, J., Hastie, T., Friedman, J., Tibshirani, R. & Simon, N. Glmnet for matlab. *Accessed: Nov* **13**, 4–2 (2013).

75. Hanks, T. D. *et al.* Distinct relationships of parietal and prefrontal cortices to evidence accumulation. *Nature* **520**, 220–223 (2015).

76. Kopec, C. D., Erlich, J. C., Brunton, B. W., Deisseroth, K. & Brody, C. D. Cortical and Subcortical Contributions to Short-Term Memory for Orienting Movements. *Neuron* **88**, 367–377 (2015).

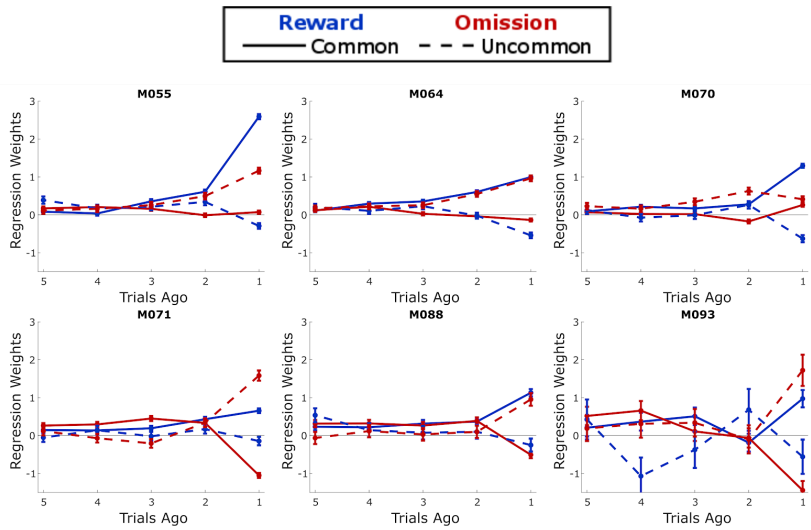

Figure S1: Weights of trial-history regression model, fit to behavioral data from the neural recording experiment. The planning strategy is characterized by a greater tendency to repeat choices that lead to rewards following a common transition than following an uncommon transition (solid blue line above dotted blue line), as well as a greater tendency to repeat choices that lead to omissions following an uncommon transition than following a common transition (dotted red line above solid red line). All six rats show evidence of such a strategy.

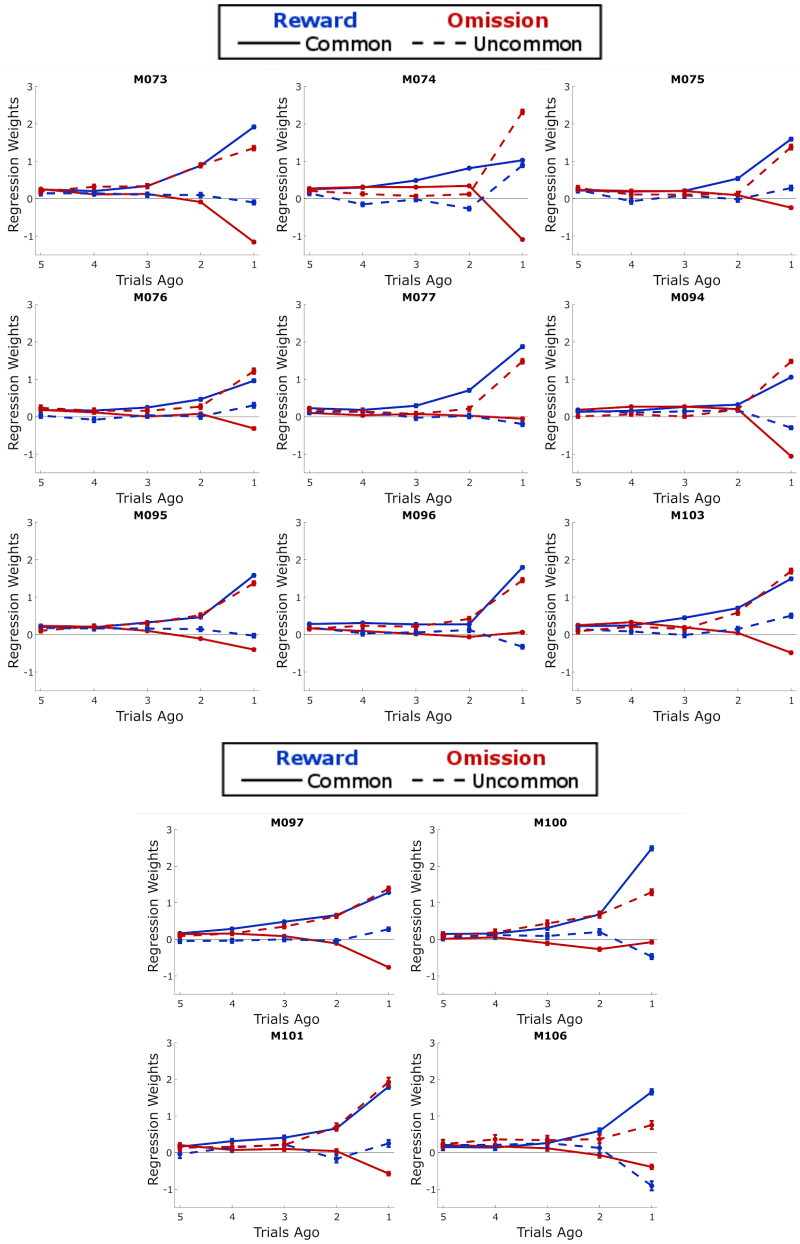

Figure S2: Weights of trial-history regression model, fit to behavioral data from the neural inactivation experiment. The planning strategy is characterized by a greater tendency to repeat choices that lead to rewards following a common transition than following an uncommon transition (solid blue line above dotted blue line), as well as a greater tendency to repeat choices that lead to omissions following an uncommon transition than following a common transition (dotted red line above solid red line). **Above:** Fits to rats implanted with optical fibers and infected with AAV-eNPH3.0. **Below:** Fits to rats implanted with optical fibers only. All rats show evidence of adopting a planning strategy.

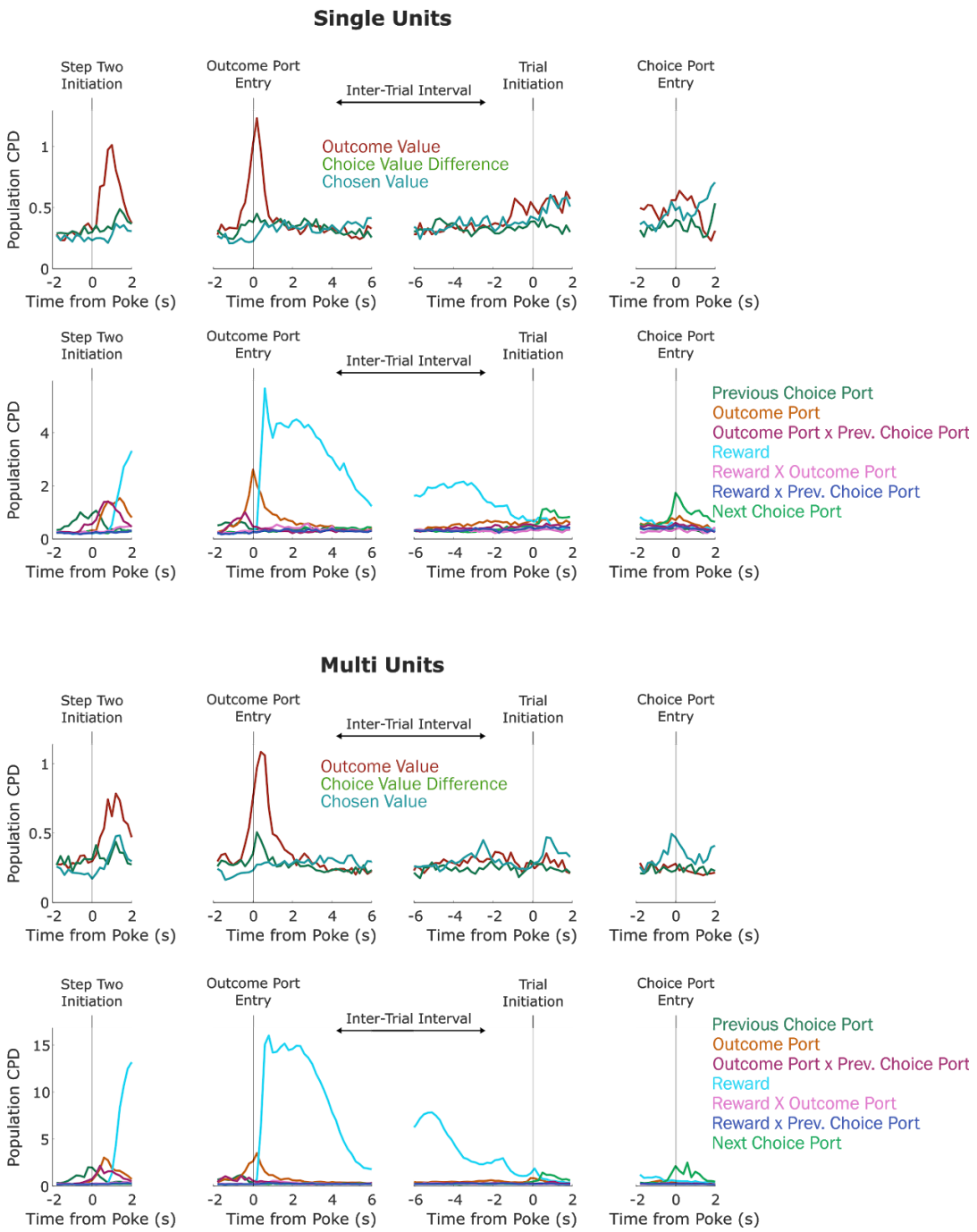

Figure S3: Timecourse of population CPD for the regressors in our model, considering only single-unit clusters (above) or considering only multi-unit clusters (below). See also Fig. 2a,c.

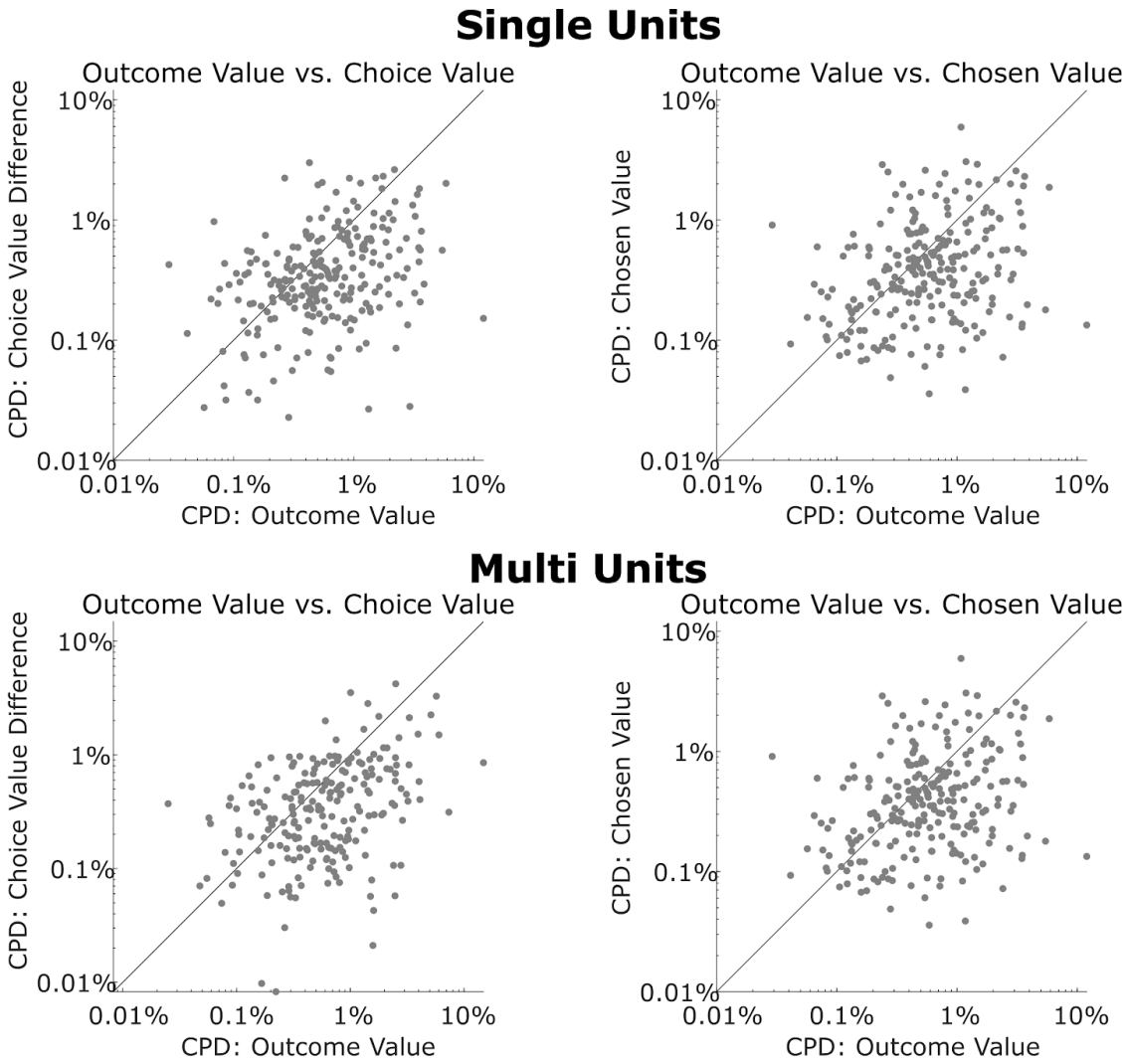

Figure S4: a. Right: Scatterplot showing CPD for each single unit (n=251) for the outcome-value regressor against CPD for the choice-value-difference regressor, both computed in a one-second window centered on entry into the outcome port. Right: Scatterplot showing CPD for outcome-value, computed at outcome port entry, against CPD for the chosen-value regressor, computed at choice port entry. **b.** As a, but for multi-unit clusters (n=226). In all panels, CPD for outcome value is greater than for choice-related value information (all $p<10^{-7}$, signrank test. See also Fig 2b).

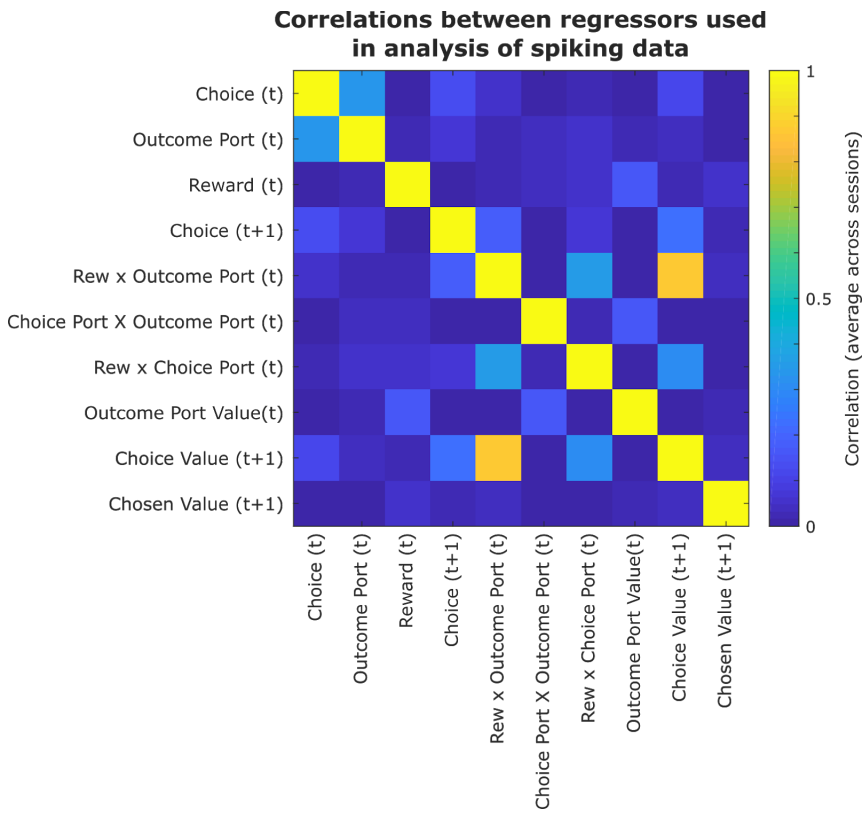

Figure S5: Correlations among predictors in the model used for analysis of electrophysiology data. Correlations among regressors require the use of a coefficient of partial determination analysis (Fig 3a, 3c) to quantify the unique contribution of each predictor to explaining neural activity.

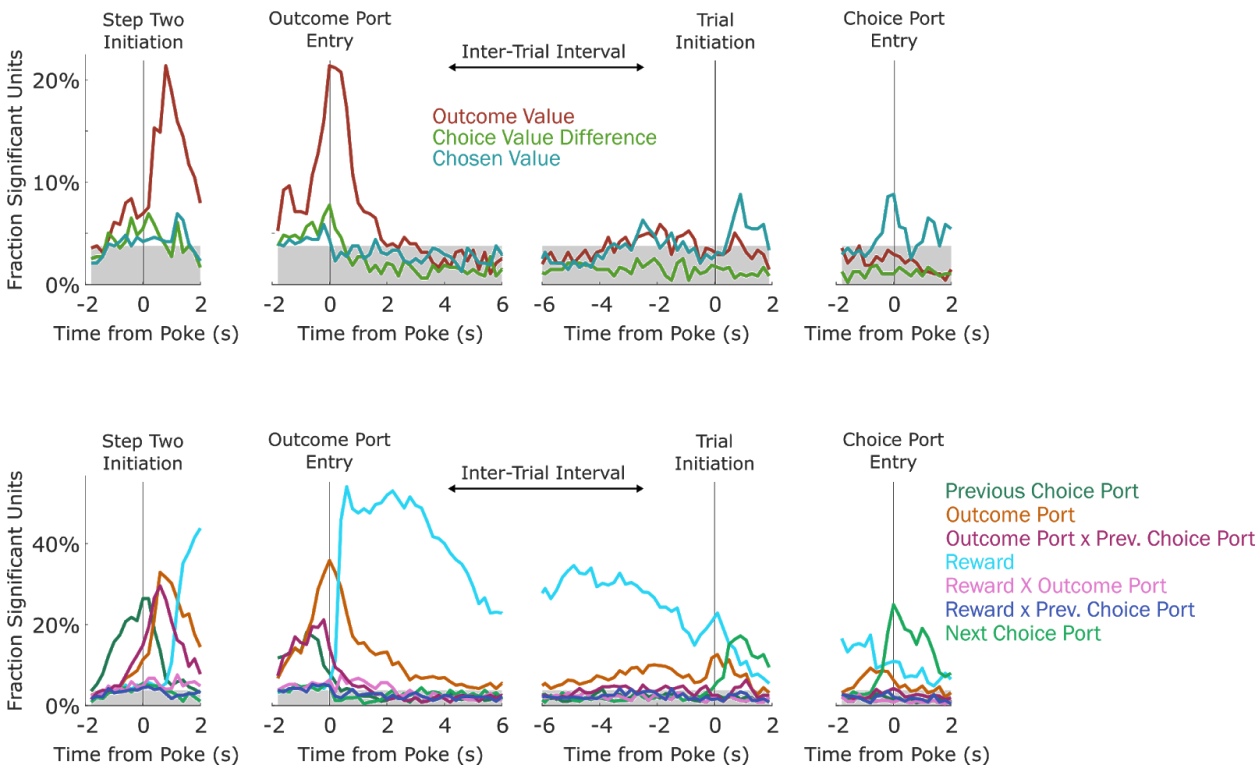

Figure S6: Fraction of units significantly encoding each predictor in each 200ms time bin. Units were deemed significant in a bin for a predictor if they earned a coefficient of partial determination larger than that of 99% of permuted datasets for that predictor in that bin. Grey shading indicates the threshold for population-wise significance (17/477 units; equivalent to $p=2x10^{-6}$ by a Bernoulli test uncorrected; $p<0.01$ after Bonferroni correction)

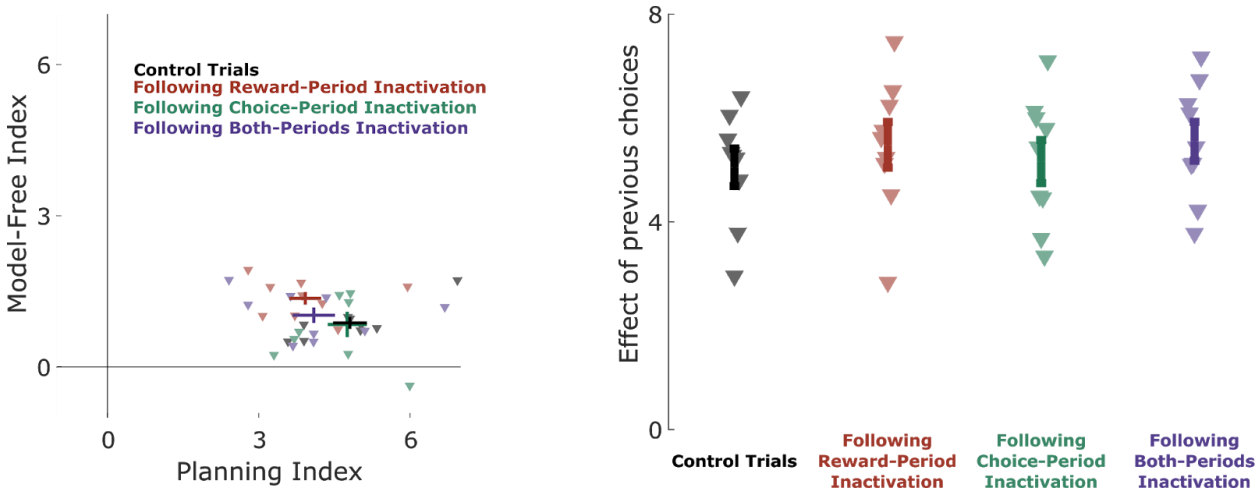

Figure S5. Left: Planning index and model-free index computed separately for control trials and for trials following optogenetic inhibition in each of the three time periods. The planning index differs significantly between control trials and trials following inhibition during either the reward period or both periods (p=0.0007, p=0.007, paired t-test; see main text for further statistics). The model-free index differs significantly between control trials and trials following reward-period inhibition (p=0.01), but not other inhibition conditions (p=0.9, p=0.5). **Right:** Total main effect of previous choices on the current choice, computed separately for control trials and for trials following optogenetic inhibition in each of the three time periods. No significant differences were found between conditions (all p> 0.05).

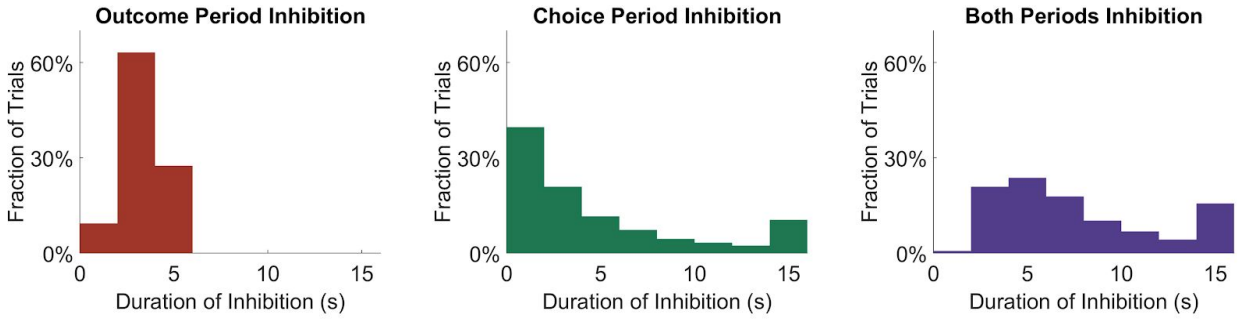

Figure S6: Duration of inhibition associated with outcome-period, choice-period, or both-periods conditions.

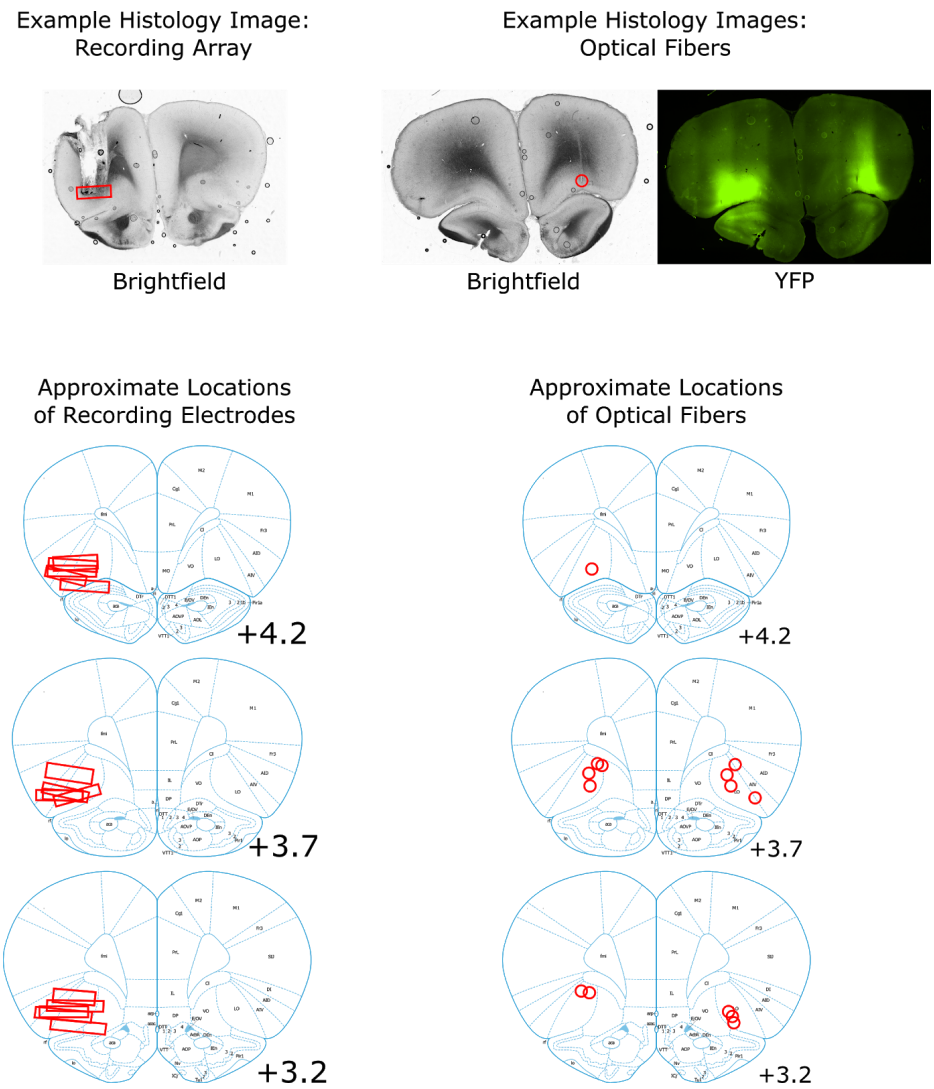

Figure S7: Histological verification of implant locations in OFC. Above left: Brightfield image of a coronal section from a rat implanted with an electrode array targeting OFC. The locations of the electrode tips are visible, as is damage done when the array was removed post-mortem. Above right: Brightfield and YFP images of a coronal section taken from a rat implanted with optical fibers and infected with AAV-halorhodopsin. The location of the optical fiber on the right is visible as a scar (optical fiber on the left is visible in a different coronal section in this rat). The area of infection is visible in the YFP channel. Below: Estimates of the locations of all recording electrodes and fiber tips, obtained by comparing histology images to the Paxinos & Watson reference atlas[52].