bioRxiv preprint doi: https://doi.org/10.1101/245720; this version posted January 17, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

Value Representations in Orbitofrontal Cortex Drive Learning, but not Choice

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

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

As humans and animals experience the world, they learn to associate states and actions with the expected values of the reward that is likely to follow[1–3]. Neural correlates of expected value are found in many brain regions, including the orbitofrontal cortex (OFC)[4–9]. While OFC value representations have been identified across many tasks and species[10–15], their computational role remains controversial[16–18]. One influential hypothesis holds that they drive value-based choosing: The OFC represents the expected values of available options, and choices are made by comparing these values to one another[4,7,9]. A contrasting hypothesis holds that they drive learning: The OFC represents the expected values of immediately impending outcomes, which are compared to rewards actually received, so as to learn and adapt expectations to match the world[5,6,19,20]. In common laboratory tasks the items to be decided between are also the items to be learned about, making the two hypothesized roles difficult to distinguish. Here, we use a recently-developed multi-step task for rats[21] that separates choosing from learning. In a first step, rats choose one of two ports ("choice ports") whose expected values are computed using planning, and are not learned. In the second step, rats are led to one of two other ports ("outcome ports") which are not chosen between, but whose expected values are learned based on reward history. We found relatively weak OFC encoding of choice port values, needed for choosing but not learning, but far stronger encoding of outcome port values, needed for learning but not choosing. Moreover, temporally-specific silencing of OFC during outcome port entry was sufficient to disrupt behavior, and the nature of this disruption was consistent with impairment of a value learning process, but was not consistent with impairment of a choice process. We therefore suggest that value representations in the OFC directly drive learning, but do not directly drive choice.

We trained rats on a two-step decision task, adapted from the human literature[22], in which a choice made by the subject in a first step is probabilistically, not deterministically, linked to an outcome that occurs in a second step (**Fig, 1a**). In each trial of our rat version of this task[21], the rat first initiated the trial by poking its nose into a neutral center port, and then made a decision between 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 tone informed the rat which of

bioRxiv preprint doi: https://doi.org/10.1101/245720; this version posted January 17, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

the two outcome ports had in fact become available on that trial, and after poking into a second neutral center poke, 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, 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 value of each outcome port *o* (which we label *R(o)*), through a learning process in which they compare their expectations to actual reward received ("Update Outcome Values" in **Fig. 1c**).

Previously, we have shown[21] that rats solve the task using a particular strategy termed "model-based planning[23–25]". This strategy utilizes an internal model of action-outcome relationships, which in this task are the probabilities linking choice ports *c* to outcome ports *o* (which we label $P(o|c)$ and which were fixed throughout the experiment, but counterbalanced across rats), to compute the expected values of the two choice ports (which we label *Q(c)*). A planning strategy results in very different roles for outcome-port values and choice-port values in our task. Outcome port values, *R(o)*, are learned incrementally from recent rewards (**Fig. 1c**, "Update Outcome Values"), while choice port values, *Q(c)*, are computed based on *R(o)* and the world model $P(o|c)$ (**Fig. 1c**, "Use World Model"**)**, and are then used to determine the next choice[21] (**Fig. 1c**, "Make Choice"). The values of the choice ports *Q(c)* therefore drive choice directly, while the learned values of the outcome ports *R(o)* support choice only indirectly. Crucially, our task features a non-deterministic $P(o|c)$, implying that $R(o_t)$ and $Q(c_t)$ will be different from one another, and can thus be estimated separately. Consistent with previous results[21], 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 $P(o|c)$ (**Fig. 1d,** planning index[21], Methods), 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, once the agent's parameters have been fit to a particular rat, the model can provide a trial-by-trial estimate of the value placed by that rat on the choice and on the outcome ports, which we compare below to trial-by-trial physiology data[13,22,26–29]. Second, the model can be altered to selectively impair information about expected values of the choice or the outcome ports, and used to generate synthetic behavioral datasets that predict the consequences of such specific impairments. Below we compare these predictions to data from rats undergoing silencing of the OFC.

bioRxiv preprint doi: https://doi.org/10.1101/245720; this version posted January 17, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

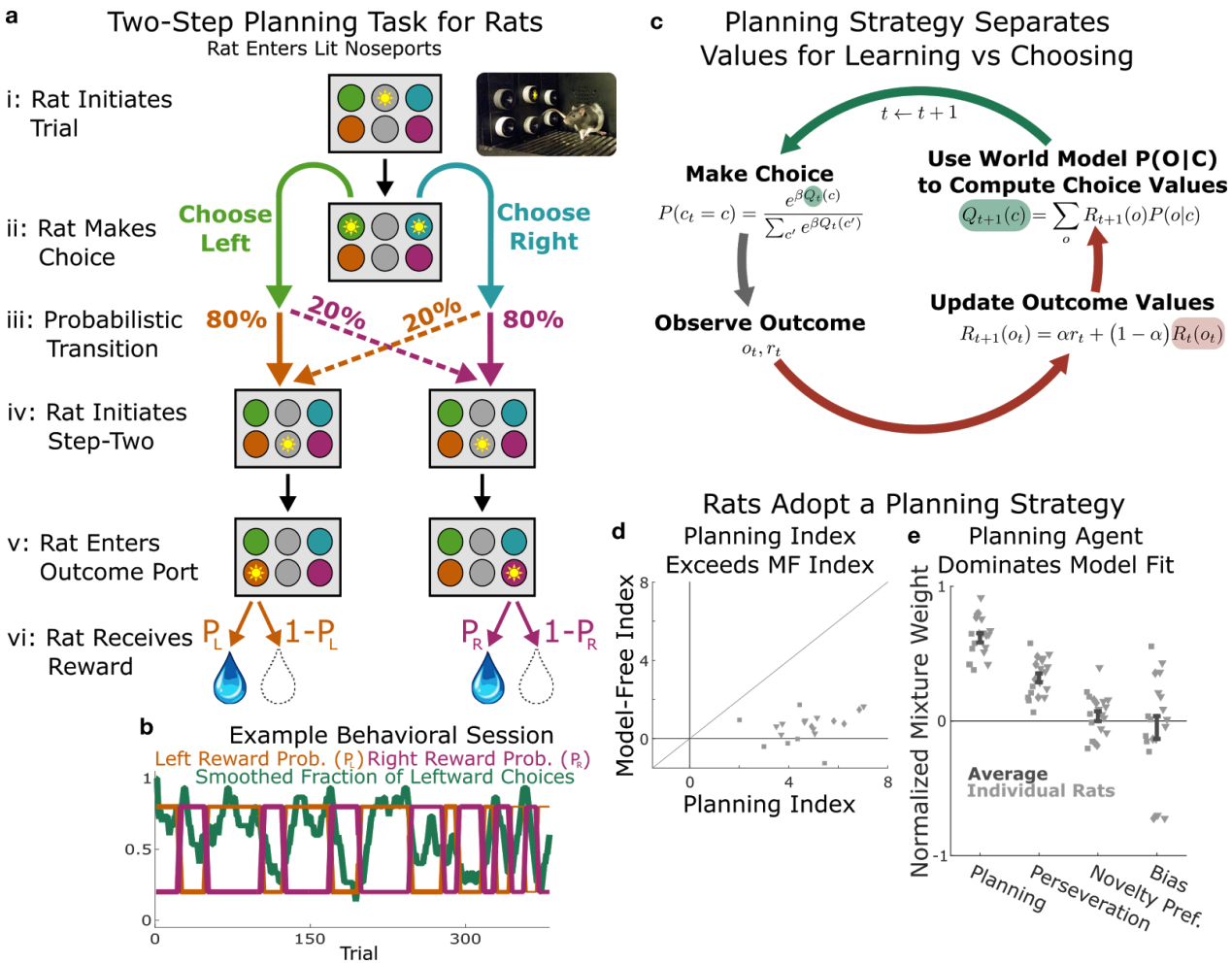

Figure 1. Two-step task separates choice values from outcome values. **a:** Two-step task optimized for rats. Rat initiates the trial by entering the lit top neutral center port (i), then indicates its decision by entering one of two lit choice ports (ii). This leads to a probabilistic transition (iii) to one of two possible paths. In both paths, rat enters the lit bottom neutral center port (v), causing one of the two outcome ports to immediately illuminate. Rat enters that outcome port (v), and receives a reward (vi). **b:** Example behavioral session. At unpredictable intervals, reward probabilities at the two outcomes port flip synchronously between high and low. Rat adjusts choices accordingly **c:** Schematic of a planning agent solving the task. Agent maintains outcome value estimates (R) for each outcome port, based on a history of recent rewards at that port, as well as choice value estimates (Q) for each choice port, which are computed on each trial based on the outcome values and the world model $P(o|c)$. **d:** Planning index and model-free index, measures that do not involve a model agent[21], shown for electrophysiology rats (n=6, squares), optogenetics rats (n=9, triangles), and sham optogenetics rats (n=4, diamonds), indicate behavior dominated by planning. The indices are calculated using a trial-history regression analysis (see *Methods* for details) **e:** Agent-based analysis: weights of an agent-based model fit to rats' behavioral data also lead to the conclusion that the behavior is dominated by planning.

Electrophysiological recordings during 51 behavioral sessions in six rats yielded 391 activity traces. Traces with an average firing rate of less than 1Hz were discarded, leaving 329 units, including both single- and multi-unit traces. Results were similar across the two (Extended Data **Fig. ED1-4**) and we therefore report both combined unless otherwise indicated. To quantify the coding properties of these units, we fit regression models to predict their spiking activity. Since the computations supporting learning and choosing must take place between outcome port entry on one trial and choice port entry on the next, we used regressors from the last few events of one

bioRxiv preprint doi: https://doi.org/10.1101/245720; this version posted January 17, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

trial (identity of the chosen choice port, identity of the outcome port rat was led to, reward presence versus omission, outcome-port-by-reward interaction, and a model-derived estimate of outcome port value[28,29], termed "outcome-value") as well as regressors from the choosing-related events of the subsequent trial (identity of port chosen in this subsequent trial, and two model-derived quantities: estimated value of this chosen port, termed "chosen-value", and difference in estimated value between the two choice ports, termed "choice-value-difference"[28,29]), and estimated to what extent spike rates depended on these regressors. We separately did this for spikes time-locked to a variety of events: we used time-locking to the last two port entry events of the first trial (bottom neutral center port; outcome port) and to the first two port entry events of the subsequent trial (top neutral center port; choice port). For each of these time-lockings, we carried out fits at each of multiple 200 ms-wide time bins around the corresponding event (Methods).

Since the different regressors are correlated with one another (Extended Data **Fig. ED6**), we quantified the contribution of each by its coefficient of partial determination (CPD; also called "partial r-squared"), which is the fraction of variance in spiking activity that regressor explained, after accounting for the influence of all other regressors[14,30]. CPD can be computed for a particular fit (i.e. one unit in one time bin), or for a collection of fits (aggregating variance over units, bins, or both). First, we considered coding in individual units, computing CPD in windows of five time bins (1s total) centered around entry into each nose port. We found that a large fraction of units significantly modulated their firing rate according to outcome-value, with the largest fraction at the time of entry into the outcome port (158/329, 48%, permutation test at $p<0.01$). In contrast, a much smaller fraction of units modulated their firing rate according to choice-value-difference, with the largest fraction at outcome port entry (34/329, 10%), or to chosen-value, with the largest fraction at choice port entry (41/329, 12%). Furthermore, the magnitude of CPD was larger for outcome-value than for the other predictors, whether taken at the port with the largest number of significant units: the median unit had CPD for outcome-value 2.1x larger than for choice-value-difference, and 2.2x larger than for chosen-value ($p=10^{-25}$, $p=10^{-32}$, signrank test; **Fig. 2a**; note logarithmic axes, and see also **Figs. ED2 ED3,** and **ED5**). Similar results were obtained by computing CPD by aggregating variance over all time bins ($p=10^{-11}$, $p=10^{-9}$, **Fig. ED1**).

Next, we considered coding at the population level, computing CPD aggregated over all units for each time bin. We found that population coding of outcome-value began to rise at entry into the neutral bottom-center port, peaking at 1.5 shortly after entry into the outcome port (**Fig 2e**). In contrast, population CPD for the two choice-related value regressors remained low in all time bins, with a maximum value of 0.51. These results indicate that neural activity in OFC encodes information about values of the outcome ports, needed for learning, much more strongly than it encodes either type of value information about the choice ports, needed for choosing.

bioRxiv preprint doi: https://doi.org/10.1101/245720; this version posted January 17, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

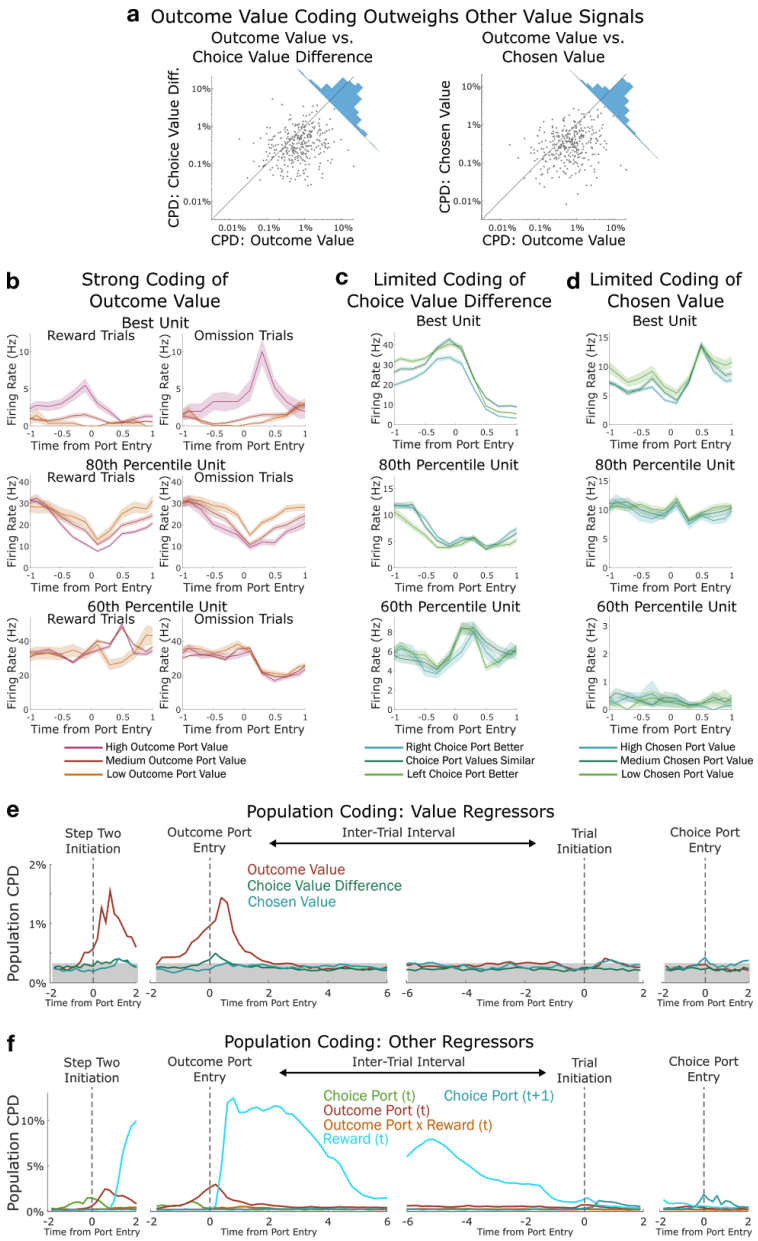

Figure 2. OFC encodes outcome values, but not choice-related values. **a**. Left: Scatterplot showing the coefficient of partial determination (CPD) for each unit (n=329) for the outcome-value regressor against CPD for the choice-value-difference regressor, both computed in a one-second window (five timebins) centered on entry into the outcome port (the port entry event with the strongest coding for both of these regressors). Right: Scatterplot showing CPD for outcome-value (computed at outcome port entry as in **a**), against CPD for the chosen-value regressor, computed at choice port entry (the port entry event with the strongest coding). **b**, Average firing rates of three example units at outcome port entry on rewarded (left) and unrewarded trials (right), separated by the expected value of the outcome port, $R_t(o_t)$. Cells displayed are the best, 80th percentile, and 60th percentile cells by reduction in sum-squared-error attributable to the outcome-value regressor (see Methods). **c**.Same as panel b, but cells are selected and separated based on the choice-value-difference regressor, $Q(left)$ - $Q(right)$. **d**. Same as panels b and c, but cells are selected and separated based on the chosen-value regressor $Q(c_t)$, and firing rates are shown at choice port entry. **e**. Timecourse of population CPD for the three value regressors. **f**. Timecourse of population CPD for the five remaining regressors.

bioRxiv preprint doi: https://doi.org/10.1101/245720; this version posted January 17, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

To help assess the causal role of the OFC's value signals, we silenced activity in the OFC using the optogenetic construct halorhodopsin (eNpHR3.0) during either the *outcome period* (beginning with 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. 3a**). Previous work[21] had shown that whole-session silencing of the OFC specifically attenuates the planning index (**Fig. 1e**), which summarizes the extent to which the rats' choices are modulated by past trials' outcomes in a way consistent with planning. Indeed, inactivation that spanned both periods decreased the planning index on the subsequent trial (p=0.007, t-test; **Fig 3b**). We found that inactivation during the outcome period alone similarly disrupted planning (p=0.0006, **Fig 3b**), in a manner indistinguishable from inactivation of both periods (p=0.47, paired t-test, for outcome period vs. both periods), and significantly greater than inactivation during the choice period (p=0.007, paired t-test). Choice period inactivation alone had no effect (p=0.65). A control group of four rats that received sham inactivation showed no effect on planning for any time period (all p>0.15; Figure 3B, grey diamonds), and significant differences were found between experimental and control rats for 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 is sufficient to disrupt planning behavior.

To assess which aspect of the behavior was affected by OFC outcome period inactivation, we perturbed our artificial planning agent in three separate ways. In the first, we attenuated choice-value representations, transiently scaling the agent's $Q_{t+1}(c)$ towards 0.5 for all c (**Fig. 3c**, left). In the second, we attenuated the reward representation, scaling r_t (**Fig. 3c**, middle). And in the third, we attenuated outcome-value representations, scaling $R_t(o_t)$ (**Fig 3c**, 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. 3c**). We therefore performed the same three past trial analysis on the experimental data **(Fig. 3d**). We 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, signrank test), but that it did affect the influence of the previous two trials' outcomes (p=0.004, p=0.04, **Fig. 3d**). Comparing to the inactivations in our artificial agent, this pattern was only reproduced by a scaling down of the outcome value $R_t(o_t)$ (compare **Fig. 3d** to **c**). This is because $R_t(o_t)$ acts as a summarized memory of the rewards of previous trials, and thus mediates the influence of past rewards (r_{t-1}, r_{t-2}, etc) on behavior (**Fig. 3e**). In contrast, scaling r_t or $Q_{t+1}(c)$ on a particular trial affects the influence of that trial's reward on the upcoming choice. We conclude that silencing the OFC in our task predominantly affects $R_t(o_t)$, values needed for learning, but has no effect on $Q_{t+1}(c)$, values needed for choosing.

bioRxiv preprint doi: https://doi.org/10.1101/245720; this version posted January 17, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

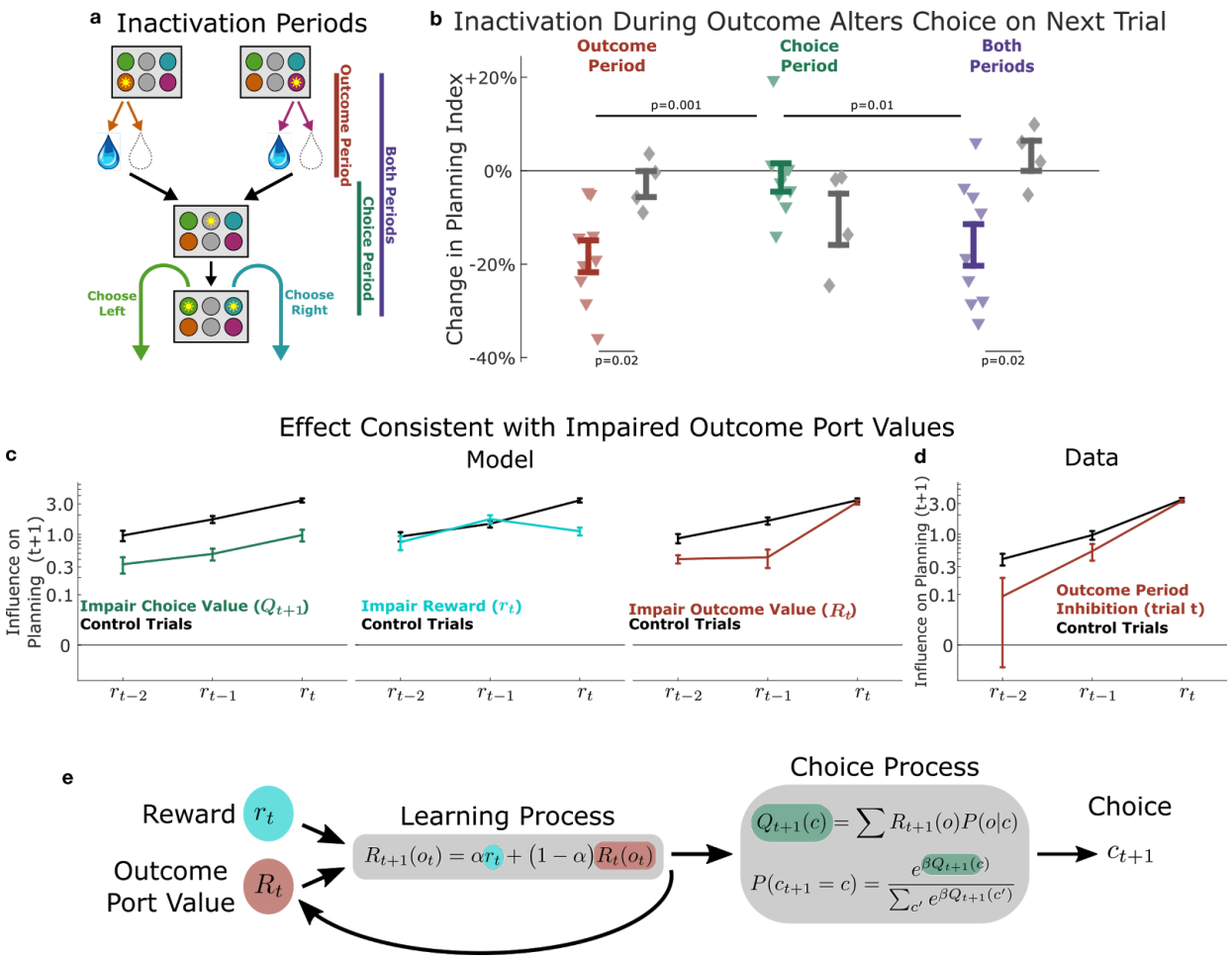

Figure 3. Inactivation of OFC attenuates influence of outcome values. **a:** Schematic showing the three inactivation time periods. Outcome-period inactivation began at the time the rat entered the outcome port, and continued until the rat exited the port, for a minimum of two seconds. Choice-period inactivation began either at the time the rat exited the outcome port or two seconds after it entered the outcome port, whichever was later, and continued until the rat entered the choice port on the next trial. Both-period inactivation encompassed both of these periods. If a scheduled inactivation would have continued for more than 15 seconds, the inactivation was terminated, and that trial was not considered for analysis. **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) in each of the three conditions. Error bars give standard errors across rats. **c:** Analysis of datasets generated with synthetic agents, illustrating different possible effects of OFC inactivation. Each panel shows the contribution to the planning index on trials ($t+1$) of rewards at different lags (r_t, r_{t-1}, r_{t-2}), both on control trials (black) and on trials with simulated inactivation on trial t (colored). Simulated inactivation consisted of either scaling the choice values (Q_{t+1}; left), scaling the reward (r_t; middle) or scaling the outcome value ($R_t(o_t)$; right) towards 0.5. Error bars give standard error across different simulated rats (see Methods) **d**: Same analysis as in c, applied to actual data from optogenetic inactivation of the OFC during the outcome period. **e**: Schematic of a model illustrating the computational roles of outcome value, choice value, and reward on choice within a model-based planning agent.

Although learning and choosing are very different reasons to compute expected value, experiments to date have not distinguished whether value signals in the OFC drive one process, the other process, or both. The two-step planning task gave us the opportunity to separate the two roles, both in terms of neural firing rate encoding, and in terms of the impact of OFC silencing.

bioRxiv preprint doi: https://doi.org/10.1101/245720; this version posted January 17, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

Both the electrophysiological and the optogenetic results challenge the influential view that the OFC directly drives choice by representing values of available options so that they can be compared[4,7,9]. We find limited representation of values associated with options, little effect of silencing OFC at the putative time of choice, and effects of silencing inconsistent with impairing choice values in a computational model. These data are consistent with the recent finding that silencing the OFC does not impair economic choice[31]. Instead, we find strong representation of values associated with immediately impending reward outcomes, a strong effect of silencing OFC at the time of those outcomes, and effects of silencing consistent with impairing outcome values in a computational model. These results thus support an alternative view, in which OFC supports choice only indirectly, by directly driving learning[5,6,19,20]. This account is consistent with recent proposals that the OFC represents a model of the task environment[32–34], or else a supports a "state space" for reinforcement learning[35–37], either of which might be used by other parts of the brain to compute the expected values of choices. While the source and detailed nature of the OFC's representations remain an important area for future research, our results resolve a key question about their computational role. Specifically, we propose that outcome value information in the OFC constitutes critical input to a learning process that updates choice values elsewhere in the brain.

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, and Jovanna Teran, Klaus Osorio, Adrian Sirko, Samantha Stein, and Lillianne Teachen for assistance with animal training. We would like to thank Yael Niv and Nathaniel Daw for helpful discussions, and Jeffrey Gauthier, Cristina Domnisoru, and Kimberly Stachenfeld for helpful comments on the manuscript.

bioRxiv preprint doi: https://doi.org/10.1101/245720; this version posted January 17, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

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[21]. 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' 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 within rat for the entirety of the animal's 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[21,38]. 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 define 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.

bioRxiv preprint doi: https://doi.org/10.1101/245720; this version posted January 17, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

We expect model-free agents to repeat choices which lead to reward and switch away from those which lead to omissions[22], 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:

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

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:

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

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

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

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[21]. 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, $Q_{total}(c)$:

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

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:

$$Q_{plan}(c) = \sum_{o} R(o) P(o|c) \tag{6}$$

bioRxiv preprint doi: https://doi.org/10.1101/245720; this version posted January 17, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

At the beginning of each session, the reward estimate *R(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

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

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**):

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

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.

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

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}.

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

Model Fitting

We implemented the model described above using the probabilistic programming language Stan[39,40], and performed maximum-a-posteriori fits using weakly informative priors on all parameters[41] 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

bioRxiv preprint doi: https://doi.org/10.1101/245720; this version posted January 17, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

procedures described in more detail elsewhere[42]. 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, on the condition that they showed an average firing rate greater than 1 Hz.

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 eight total regressors, defined relative to pairs of adjacent trials, and consisting of the choice port selected (left or right), the outcome port visited (left or right), the reward received (reward or omission), the interaction between outcome port and reward, and the expected value of the outcome port visited (*V*) for the first trial, and the choice port selected, the value difference between the choice ports (*Q(left)* - *Q(right)*), and the value of the choice port selected (*Q(chosen)*) for 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 `lassoglm` using a Poisson noise model and L1 regularization parameters (pure lasso regression; $\alpha = 1$, $\lambda = 10^{-4}$) sufficient to yield a non-null model for all units.
In our task, many of these regressors were correlated with one another (**Fig. ED6**), so we quantify encoding using the coefficient of partial determination (CPD; also known as partial r-squared) associated with each[14,30]. 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 each unit taken over all time bins (Fig ED1), 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 permuted 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 sign test on the differences in CPD.

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

bioRxiv preprint doi: https://doi.org/10.1101/245720; this version posted January 17, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

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

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[42–44], 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, inactivation sessions began. During these sessions, the laser shutter was opened (causing light to flow into the implant, activate 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.

bioRxiv preprint doi: https://doi.org/10.1101/245720; this version posted January 17, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

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 inactivation is most likely to have no effect on behavior, and than any effect is has is equally likely to be positive or negative with respect to each β. 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 signrank test across rats.

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:

bioRxiv preprint doi: https://doi.org/10.1101/245720; this version posted January 17, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

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

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. Sharp optical fibers do not leave a visible scar near their tips, so we estimated the position of the tips using the trajectory of the scar and the known distance below brain surface to which fibers were lowered during surgery. Location of virus expression was identified by imaging the GFP conjugated to the eNpHR3.0 molecule. Note that at the time of the first submission of this paper, histological verification of targeting is still underway.

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 reasonable request. Software used for training rats is available on the Brody lab website.

bioRxiv preprint doi: https://doi.org/10.1101/245720; this version posted January 17, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

Citations

1. Lee, D., Seo, H. & Jung, M. W. Neural basis of reinforcement learning and decision making. *Annu. Rev. Neurosci.* **35,** 287–308 (2012).
2. 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).
3. 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).
4. Wallis, J. D. Orbitofrontal cortex and its contribution to decision-making. *Annu. Rev. Neurosci.* (2007).
5. 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).
6. 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).
7. Padoa-Schioppa, C. & Conen, K. E. Orbitofrontal Cortex: A Neural Circuit for Economic Decisions. *Neuron* **96,** 736–754 (2017).
8. Kringelbach, M. L. The human orbitofrontal cortex: linking reward to hedonic experience. *Nat. Rev. Neurosci.* **6,** 691–702 (2005).
9. Padoa-Schioppa, C. Neurobiology of economic choice: a good-based model. *Annu. Rev. Neurosci.* **34,** 333–359 (2011).
10. Gallagher, M., McMahan, R. W. & Schoenbaum, G. Orbitofrontal cortex and representation of incentive value in associative learning. *J. Neurosci.* **19,** 6610–6614 (1999).
11. Plassmann, H., O'Doherty, J. & Rangel, A. Orbitofrontal cortex encodes willingness to pay in everyday economic transactions. *J. Neurosci.* **27,** 9984–9988 (2007).
12. Padoa-Schioppa, C. & Assad, J. A. Neurons in the orbitofrontal cortex encode economic value. *Nature* **441,** 223–226 (2006).
13. 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).
14. 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).
15. Gottfried, J. A., O'Doherty, J. & Dolan, R. J. Encoding predictive reward value in human amygdala and orbitofrontal cortex. *Science* **301,** 1104–1107 (2003).
16. Stalnaker, T. A., Cooch, N. K. & Schoenbaum, G. What the orbitofrontal cortex does not do. *Nat. Neurosci.* **18,** 620–627 (2015).
17. Murray, E. A., O'Doherty, J. P. & Schoenbaum, G. What we know and do not know about the functions of the orbitofrontal cortex after 20 years of cross-species studies. *J. Neurosci.* **27,** 8166–8169 (2007).
18. Wallis, J. D. Cross-species studies of orbitofrontal cortex and value-based decision-making. *Nat. Neurosci.* **15,** 13–19 (2012).
19. McDannald, M. A. *et al.* Model-based learning and the contribution of the orbitofrontal cortex to the model-free world. *Eur. J. Neurosci.* **35,** 991–996 (2012).
20. Takahashi, Y. K. *et al.* The orbitofrontal cortex and ventral tegmental area are necessary for learning from unexpected outcomes. *Neuron* **62,** 269–280 (2009).
21. Miller, K. J., Botvinick, M. M. & Brody, C. D. Dorsal hippocampus contributes to model-based planning. *Nat. Neurosci.* **20,** 1269–1276 (2017).
22. 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).

bioRxiv preprint doi: https://doi.org/10.1101/245720; this version posted January 17, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

23. 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).
24. Dolan, R. J. & Dayan, P. Goals and habits in the brain. *Neuron* **80,** 312–325 (2013).
25. Sutton, R. S. & Barto, A. G. *Reinforcement learning: An introduction*. **1,** (MIT press Cambridge, 1998).
26. Barraclough, D. J., Conroy, M. L. & Lee, D. Prefrontal cortex and decision making in a mixed-strategy game. *Nat. Neurosci.* **7,** 404–410 (2004).
27. Sugrue, L. P., Corrado, G. S. & Newsome, W. T. Matching behavior and the representation of value in the parietal cortex. *Science* **304,** 1782–1787 (2004).
28. Corrado, G. & Doya, K. Understanding neural coding through the model-based analysis of decision making. *J. Neurosci.* **27,** 8178–8180 (2007).
29. Daw, N. D. Trial-by-trial data analysis using computational models. in *Decision Making, Affect, and Learning* 3–38 (2011).
30. 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).
31. 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).
32. Stalnaker, T. A. *et al.* Orbitofrontal neurons infer the value and identity of predicted outcomes. *Nat. Commun.* **5,** 3926 (2014).
33. McDannald, M. A., Esber, G. R., Wegener, M. A. & Wied, H. M. Orbitofrontal neurons acquire responses to'valueless' Pavlovian cues during unblocking. *Elife* (2014).
34. 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).
35. Wilson, R. C., Takahashi, Y. K., Schoenbaum, G. & Niv, Y. Orbitofrontal cortex as a cognitive map of task space. *Neuron* **81,** 267–279 (2014).
36. Schuck, N., Wilson, R. & Niv, Y. A state representation for reinforcement learning and decision-making in the orbitofrontal cortex. *bioRxiv* (2017). doi:10.1101/210591
37. Takahashi, Y. K., Stalnaker, T. A., Roesch, M. R. & Schoenbaum, G. Effects of inference on dopaminergic prediction errors depend on orbitofrontal processing. *Behav. Neurosci.* **131,** 127–134 (2017).
38. Lau, B. & Glimcher, P. W. Dynamic response-by-response models of matching behavior in rhesus monkeys. *J. Exp. Anal. Behav.* **84,** 555–579 (2005).
39. Stan Development Team. MatlabStan: The MATLAB interface to Stan. (2016).
40. 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).
41. Gelman, A. *et al. Bayesian Data Analysis, Third Edition*. (CRC Press, 2013).
42. 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
43. Hanks, T. D. *et al.* Distinct relationships of parietal and prefrontal cortices to evidence accumulation. *Nature* **520,** 220–223 (2015).
44. 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).

bioRxiv preprint doi: https://doi.org/10.1101/245720; this version posted January 17, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

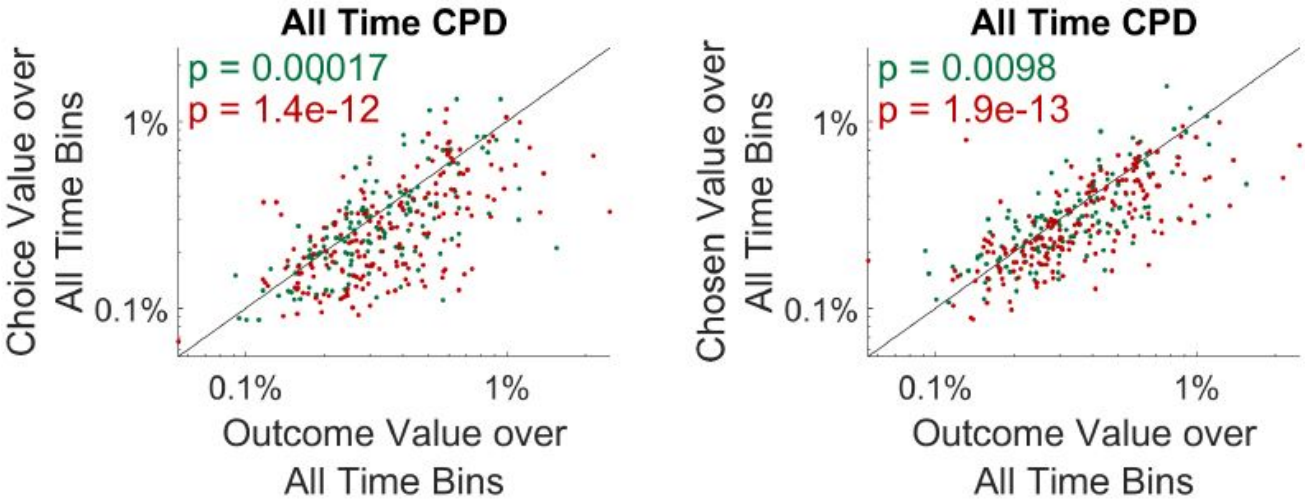

Figure ED1: Left: Coefficient of Partial Determination (CPD) for the outcome-value and the choice-value-difference regressors, computed aggregating variance over all time bins for each single unit (green) and multi-unit (red) cluster. P-values shown are the result of a sign test over units. Right: CPD for the outcome-value and the chosen-value predictors.

bioRxiv preprint doi: https://doi.org/10.1101/245720; this version posted January 17, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

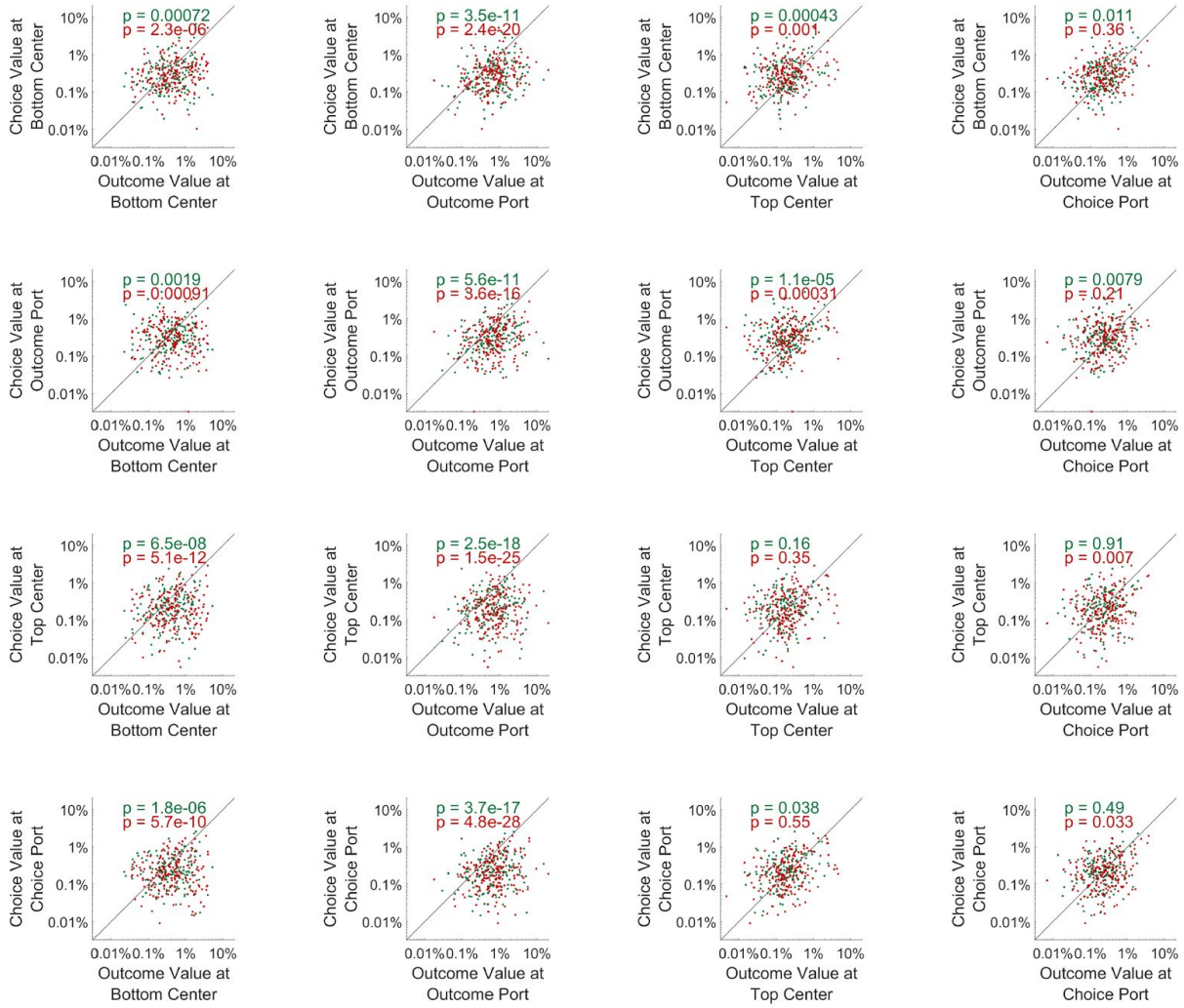

Figure ED2: Coding of outcome value and choice value difference at the time of port entry. Each panel shows CPD for the outcome-value and the choice-value-difference regressor,each computed in a one-second window (five 200ms time bins) centered on a different port entry event. P-values shown are the result of a sign test over units.

bioRxiv preprint doi: https://doi.org/10.1101/245720; this version posted January 17, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

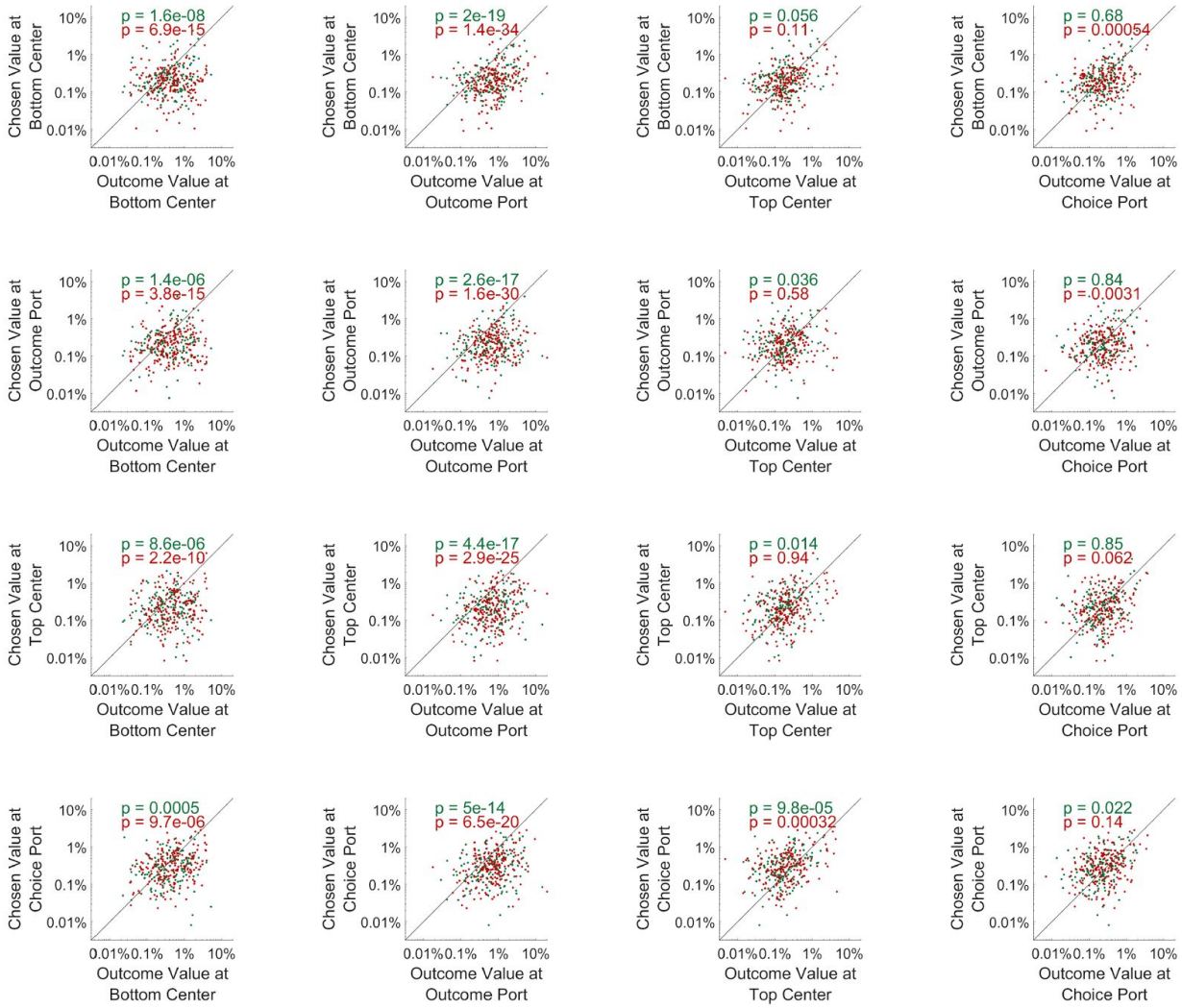

Figure ED3: Coding of outcome value and chosen value at the time of port entry. Each panel shows CPD for the outcome-value and the choice-value-difference regressor, each computed in a one-second window (five 200ms time bins) centered on a different port entry event for single-unit (green) and multi-unit (red) clusters. P-values shown are the result of a sign test over units.

bioRxiv preprint doi: https://doi.org/10.1101/245720; this version posted January 17, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

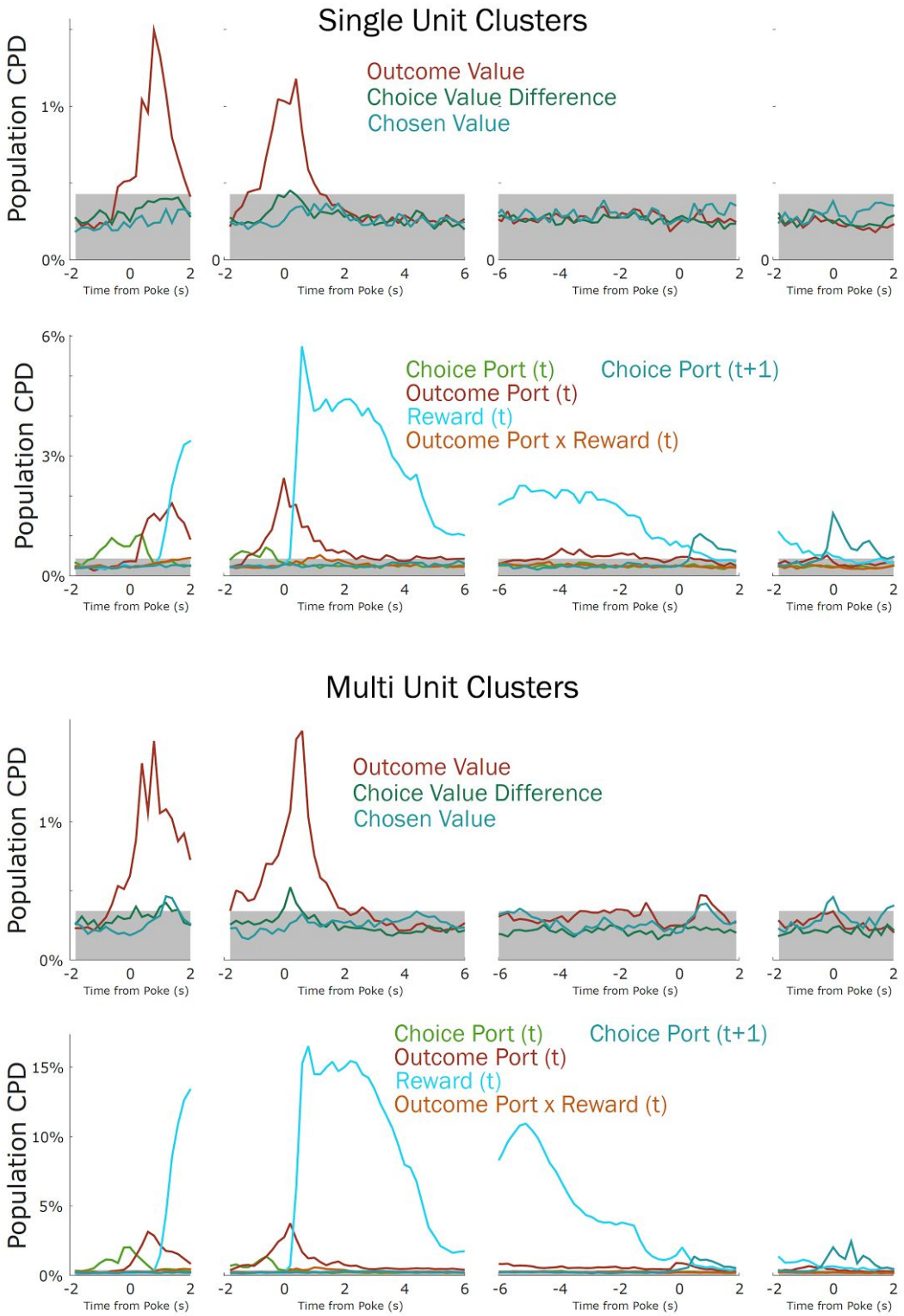

Figure ED4: Timecourse of population CPD for the six predictors in our model (see also **Fig. 2ef**), considering only single-unit clusters (above) or considering only multi-unit clusters (below)

bioRxiv preprint doi: https://doi.org/10.1101/245720; this version posted January 17, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

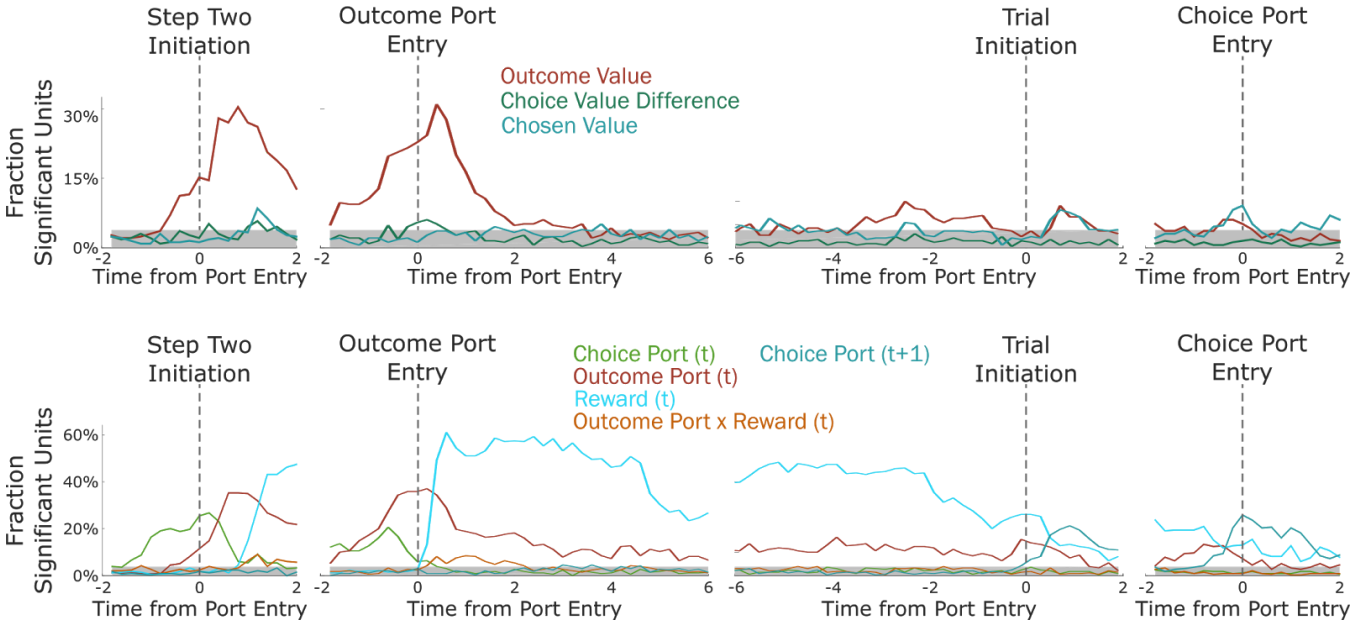

Figure ED5: Fraction of units significantly encoding each predictor in each 200ms time bin. Units were deemed significant in a bin for a predictor if they earned a coefficient of partial determination larger than that of 99% of permuted datasets for that predictor in that bin. This plot includes both single- and multi-unit clusters.

bioRxiv preprint doi: https://doi.org/10.1101/245720; this version posted January 17, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

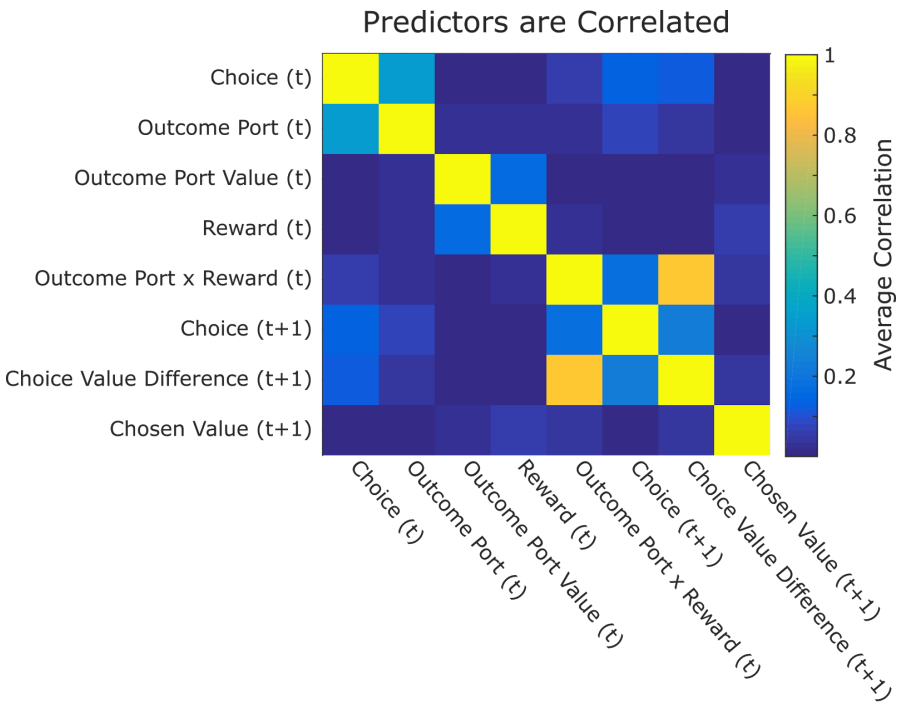

Figure ED6: Correlations among predictors in the model

bioRxiv preprint doi: https://doi.org/10.1101/245720; this version posted January 17, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

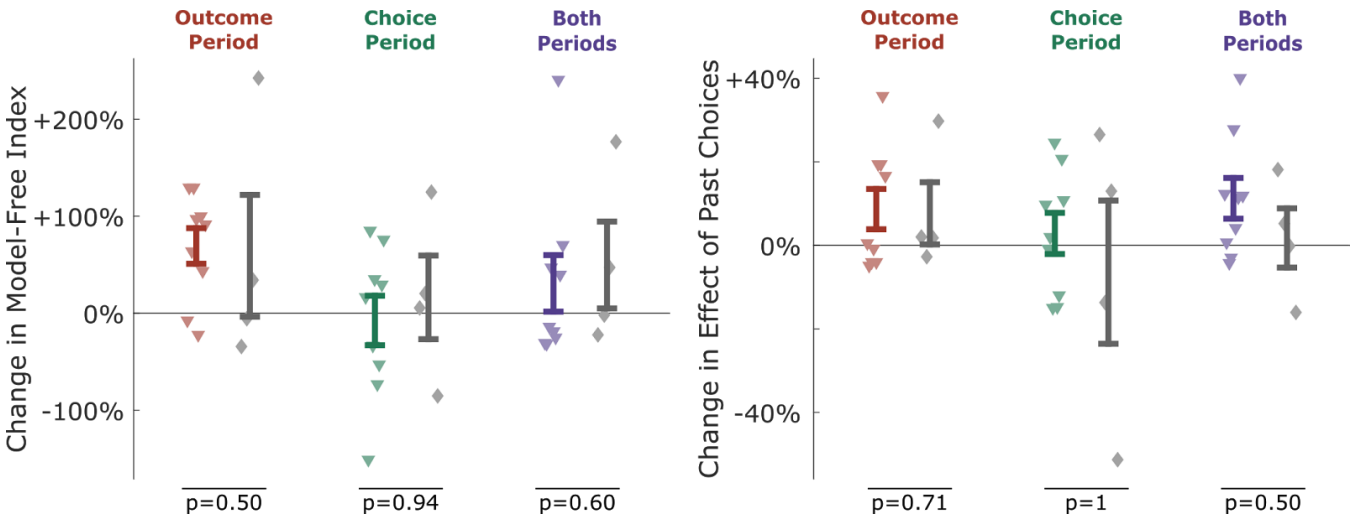

Figure ED7: Effects of optogenetic inhibition on the model-free index and on the main effect of past choices. No significant differences were found between inhibition and sham inhibition on either of these measures (rank sum tests, all $p > 0.5$).

bioRxiv preprint doi: https://doi.org/10.1101/245720; this version posted January 17, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

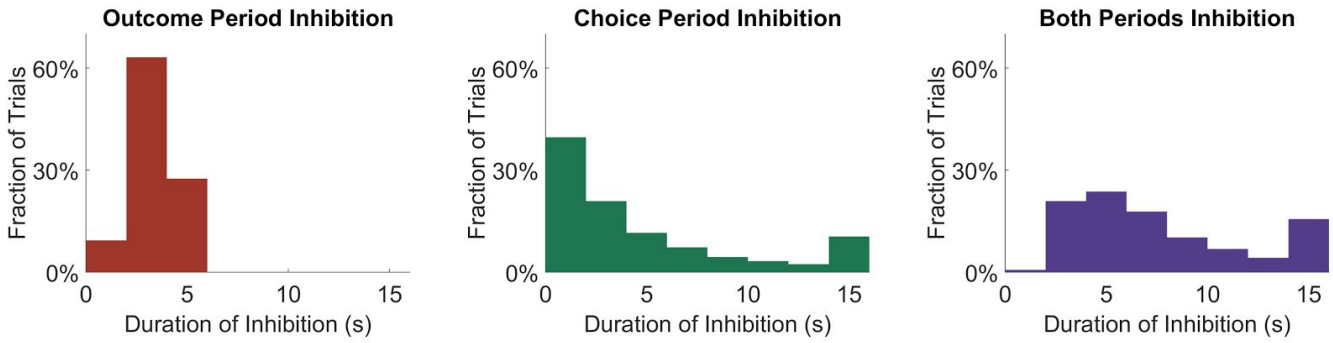

Figure ED8: Duration of inhibition associated with outcome-period, choice-period, or both-periods conditions.