bioRxiv preprint doi: https://doi.org/10.1101/2022.01.25.477716; this version posted January 26, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity.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-NC-ND 4.0 International license.

The role of the orbitofrontal cortex in creating cognitive maps

Kauê Machado Costa[1*], Robert Scholz[2,3], Kevin Lloyd[2], Perla Moreno-Castilla[4], Matthew P. H. Gardner[5], Peter Dayan[2,6] and Geoffrey Schoenbaum[1*]

[1] National Institute on Drug Abuse Intramural Research Program, National Institutes of Health, Baltimore, MD, 21224, USA

[2] Max Planck Institute for Biological Cybernetics, Tübingen, 72076, Germany

[3] Max Planck School of Cognition, Leipzig, 04103, Germany

[4] National Institute on Aging Intramural Research Program, National Institutes of Health, Baltimore, MD, 21224, USA

[5] Concordia University, Montreal, QC, 7141, Canada

[6] University of Tübingen, Tübingen, 72074, Germany

* Corresponding authors: kaue.m.costa@gmail.com and geoffrey.schoenbaum@nih.gov

Abstract

We use internal models of the external world to guide behavior, but little is known about how these cognitive maps are *created*. The orbitofrontal cortex (OFC) is typically thought to access these maps to support model-based decision-making, but it has recently been proposed that its critical contribution may be instead to integrate information into existing and new models. We tested between these alternatives using an outcome-specific devaluation task and a high-potency chemogenetic approach. We found that selectively inactivating OFC principal neurons when rats learned distinct cue-outcome associations, but prior to outcome devaluation, disrupted subsequent model-based inference, confirming that the OFC is critical for creating new cognitive maps. However, OFC inactivation surprisingly led to generalized devaluation. Using a novel reinforcement learning framework, we demonstrate that this effect is best explained not by a switch to a model-free system, as would be traditionally assumed, but rather by a circumscribed deficit in defining credit assignment precision during model construction. We conclude that the critical contribution of the OFC to learning is regulating the specificity of associations that comprise cognitive maps.

One Sentence Summary: OFC inactivation impairs learning of new specific cue-outcome associations without disrupting model-based learning in general.

Keywords: inference, chemogenetics, reinforcement learning, learning theory, model-based learning, JHU37160, cognitive map, devaluation, conditioned taste aversion.

bioRxiv preprint doi: https://doi.org/10.1101/2022.01.25.477716; this version posted January 26, 2022. 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-NC-ND 4.0 International license.

Introduction

Animals behave in ways that suggest that the brain can build, store, and use internal representations that account for the predictive relationships between elements in the external world. Also called associative models or cognitive maps[1], these mental constructs are thought to be especially important for adaptive behavior under new or changed conditions[2,3]. The inability to use such models properly is thought to be a key feature of mental illnesses such as schizophrenia[4], substance use disorder[5,6], and obsessive compulsive disorder[7]. However, despite their importance, we are only beginning to understand the informational structure of cognitive maps and how the brain creates, stores, and uses them.

In this regard, the orbitofrontal cortex (OFC) has been extensively implicated in model-based behaviors [8–11]. However, its exact contributions to defining or using the cognitive maps that support these behaviors are still controversial. The currently prevailing view is that the OFC accesses information stored elsewhere to represent the current task space at the time a decision is made [12–15]. While broadly consistent with the literature, this view is most strongly supported by devaluation experiments in which pairing a given outcome with illness (or satiety) leads to reduced conditioned responding to a cue predicting that outcome. This effect has been shown repeatedly and across species to require the OFC at the time of the probe test [16–19], a result generally interpreted as showing a role for OFC in using the associative map formed earlier in training. Compromising the OFC disrupts this usage, resulting in supposedly “model-free” or habit-like behavior. By this account, the OFC offers a form of specialized working memory required for mental simulation.

However, recent studies suggest that the OFC might instead serve as the cognitive “cartographer”, playing a critical role not merely in using maps drawn by other regions but rather in creating and modifying the maps on which other regions rely [20]. According to this view, OFC manipulations in devaluation probes affect behavior not because OFC is required for mental simulation but rather because the test requires changes to, or recombinations of, existing cognitive maps.

A logical, but untested, corollary of this alternative proposal is that the OFC should also be necessary during initial conditioning in the reinforcement devaluation task, when a major part of the cognitive map used in the later probe is *created*. On the other hand, if the classic view is correct – that, at the time of decision-making, the OFC just uses maps made and maintained elsewhere – then this region should not be necessary during the conditioning phase. As one cannot read what is not yet written, this prediction allows for an acid test to differentiate whether the OFC is a reader or a cartographer of cognitive maps. Here, we performed this test using a within-subject outcome-specific devaluation task and high-potency chemogenetics to inactivate OFC transiently when maps were first being formed.

bioRxiv preprint doi: https://doi.org/10.1101/2022.01.25.477716; this version posted January 26, 2022. 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-NC-ND 4.0 International license.

Results

Food restricted rats, transfected with either hM4d (inhibitory DREADD receptor, n=15) or only mCherry (control; n=13) in the OFC (Figure S1), underwent conditioning in which two different auditory cues (A and B) predicted the delivery of either banana- or bacon-flavored pellets (Figure 1A). Before each session, rats were injected with JHU37160 dihydrochloride (JH60; i.p. 0.2 mg/kg), a high-potency DREADD agonist [21], to inactivate OFC principal neurons in hM4d-transfected rats both transiently and selectively [22]. The use of this new generation compound avoids several confounds associated with other DREADD agonists [21,23].

Despite inactivation, rats in both groups progressively increased responding to the food cup during presentation of either cue (Figure 1D). Initial acquisition rates were similar, although rats in the hM4d group responded slightly less during the last two sessions of conditioning, in agreement with recent work showing that transient OFC inactivation can reduce asymptotic conditioned responding in some settings [24].

After conditioning, rats were subjected to conditioned taste aversion (CTA) training, in which one of the rewards (the one associated with B), was paired with LiCl injections, inducing nausea (Figure 1B). Rats initially preferred both rewards equally, but quickly and selectively reduced consumption of the pellet type paired with LiCl (Figure 1E).

Finally, after CTA training, rats were given a probe test, in which the cues were presented as during conditioning but without reward (Figure 1C). As expected, control rats responded more to cue A (paired with the non-devalued pellet) than to cue B (paired with the devalued pellet), indicating they had learned the specific cue-reward and reward-illness associations and were able to integrate them in the probe test to infer that B might lead to devalued reward (Figure 1F). By contrast, rats in the hM4d group responded equally to both cues (Figure 1F). This result is inconsistent with the hypothesis that OFC's main function is to access mental maps stored elsewhere to support model-based behaviors at the time a decision is made, and instead supports the alternative hypothesis that OFC plays a critical role in drawing those maps during initial learning [20].

That said, while this result supports this alternative hypothesis, rats in the hM4d group did not simply lack the devaluation effect, as would be expected if there was no model, but rather they appeared to generalize the devaluation effect across cues (Figure 1F). This was evident even if responses during the probe were normalized to the end of conditioning, indicating that the effect was not related to modest reduction in asymptotic conditioned responding (Figure 2A). That the two effects were orthogonal to each other is further supported by the lack of correlation between responding at the end of conditioning and the effect of devaluation (Figure 2B). Nor was the apparent generalization due to differences in CTA retention as preference tests revealed that CTA effects were similar in the two groups after the probe test (Figure 2C).

bioRxiv preprint doi: https://doi.org/10.1101/2022.01.25.477716; this version posted January 26, 2022. 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-NC-ND 4.0 International license.

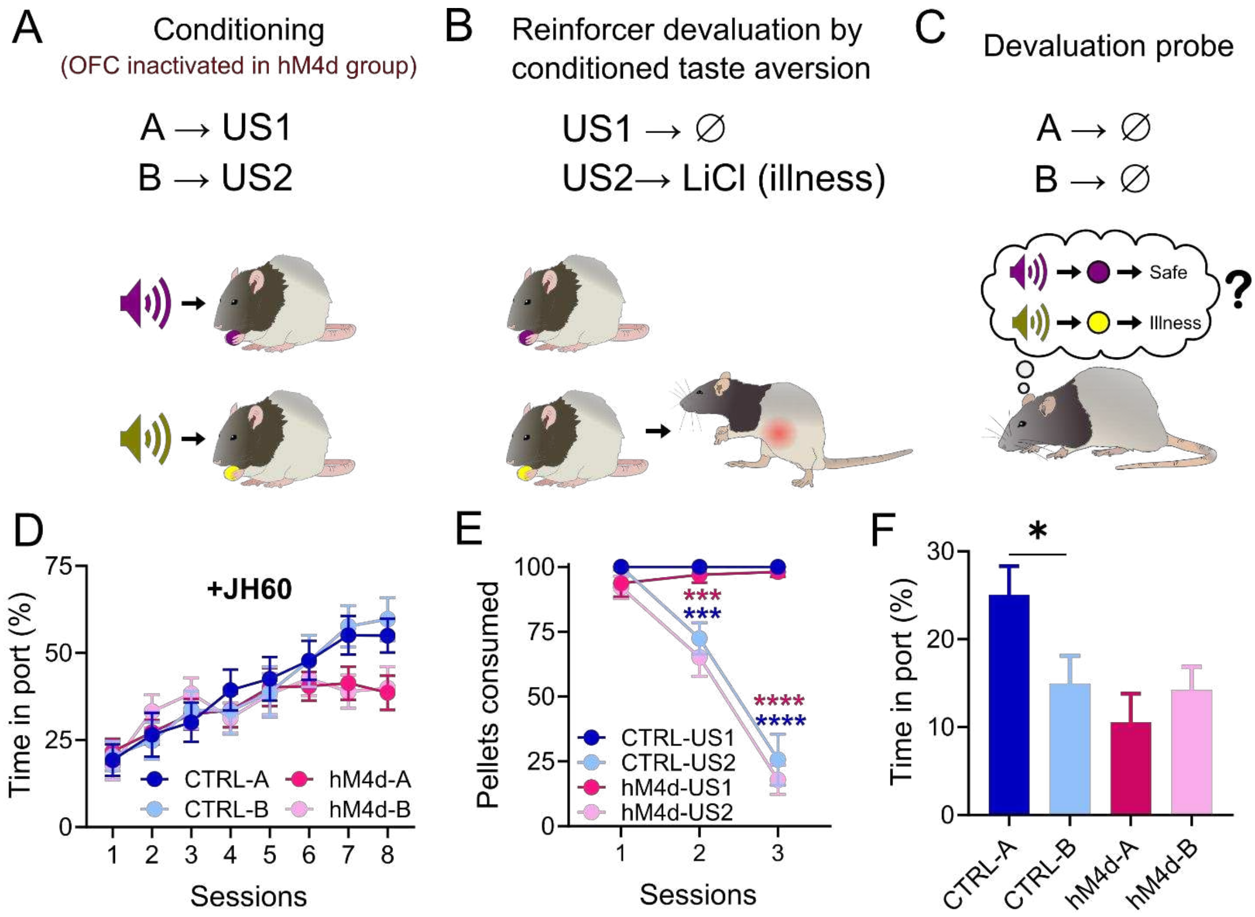

Figure 1. Chemogenetic inactivation of OFC during conditioning abolishes subsequent sensory-specific cue devaluation. (A-C): Schematic of the behavioral procedures. **(A):** Rats were conditioned to two cues, A and B, which lead to different rewards. The OFC was inactivated in the hM4d group. **(B):** Later, one of the rewards was paired with LiCl injections. **(C):** Finally, rats were re-exposed to the conditioned cues, testing if a model-based association had been established between them and the rewards. **(D):** Food cup responding during conditioning. There was no isolated or interaction effect of cue identity (*P*>0.05), nor an effect of group (*P*>0.05), and rats of both groups increased responding as sessions progressed (*P*<0.0001****). However, there was a significant interaction between group and session progression (*P*<0.0001****), visible in the last two sessions. **(E):** Pellet consumption during CTA. Rats from both groups consumed nearly all pellets in the first CTA session, and consumed less of the devalued pellet type as sessions progressed (*P*<0.0001****). **(F):** Food cup responding during probe. There was a significant effect of group (*P*=0.047*), and the interaction of the group with the cues (*P*=0.009**), as control rats responded more to A than to B, while hM4d rats responded equally to both cues. Asterisks in graphs indicate post-hoc multiple comparison test results. See Table S1 for detailed statistics. Data are represented as mean ± SEM. **P*<0.05; ****P*<0.001; *****P*<0.0001.

bioRxiv preprint doi: https://doi.org/10.1101/2022.01.25.477716; this version posted January 26, 2022. 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-NC-ND 4.0 International license.

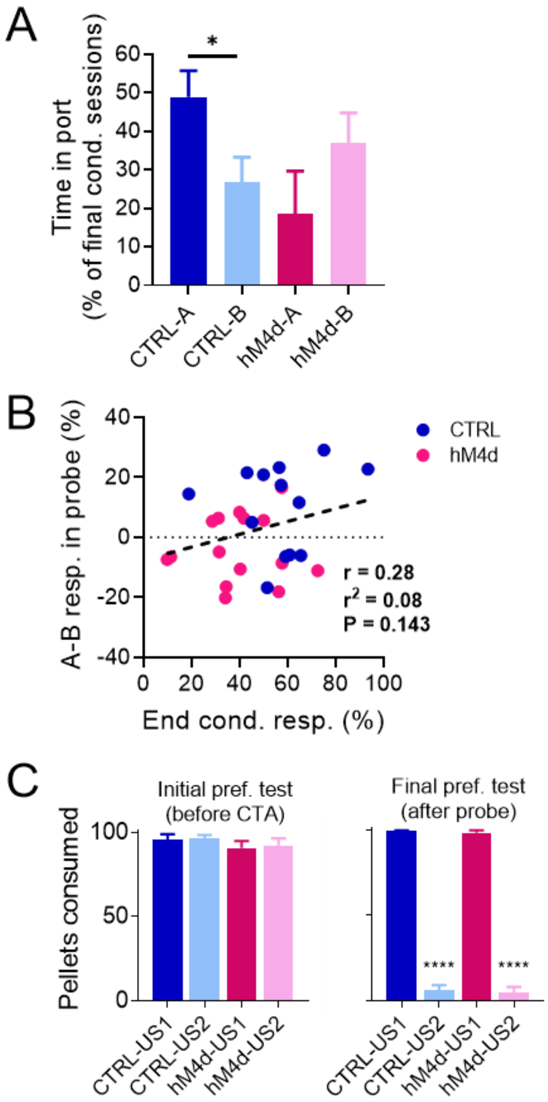

Figure 2. Generalization of devaluation due to OFC inactivation is not dependent on effects on conditioned responding or CTA learning. (A): Food port responding in the final probe session but normalized to the last two sessions of conditioning. There was a significant interaction effect of the group with the cues (*P* = 0.002**), as well as only a significant difference between A and B in the control group in the post-hoc test. **(B):** Differential responding to valued and devalued cues (responding to A – mean responding to B) was not correlated to the conditioned responding at the end of initial learning (average of % time in port for both cues in the last two sessions of conditioning). **(C):** Consumption of pellets during preference tests for CTRL (blue and light blue) and hM4d (red and pink) rats. Rats from both groups consumed all pellets similarly during the first preference test (2-way ANOVA; ND x D: $F_{1,26}$ = 0.12, *P* = 0.7318; CTRL vs hM4d: $F_{1,26}$ = 1.235, *P* = 0.2766; interaction: $F_{1,26}$ = 0.0171, *P* = 0.8969) and both groups equally consumed significantly less of the devalued pellet type (the one previously associated with cue B and paired with LiCl during CTA) in the second preference test (2-way ANOVA; ND x D: $F_{1,26}$ = 1364, *P* < 0.0001****; CTRL vs hM4d: $F_{1,26}$ = 0.3519, *P* = 0.5582; interaction: $F_{1,26}$ = 0.0005, *P* = 0.9825). Asterisks in the graphs indicate results of post-hoc multiple comparison tests. Data are represented as mean ± SEM. ****P<0.0001.

bioRxiv preprint doi: https://doi.org/10.1101/2022.01.25.477716; this version posted January 26, 2022. 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-NC-ND 4.0 International license.

Generalization of devaluation also could not be accounted for by effects of OFC inactivation on perception or memory. To show this, we tested a subset of these rats in an object recognition task [25]. OFC was inactivated prior to the sample phase of the task, while the rats first explored two identical objects (Figure 3A). Over the next 2 days, the rats were brought back to the same arena for two recognition tests in which novel objects were substituted for the objects introduced in the sample phase (Figure 3B-C). If OFC inactivation in these rats induced perceptual confusion, accelerated forgetting, or context-dependent learning, then inactivation in the sample phase of this task should have disrupted object discrimination in the first but not the second recognition test, yet we found no such effect (Figure 3D-I).

The generalization of devaluation in the OFC inactivated group was unexpected and intriguing, since model-based learning is traditionally treated as an all-or-none phenomenon. A complete failure of model-based control would leave only devaluation-insensitive, model-free behavior intact, resulting in high responding to both cues. It has been proposed that associative learning may operate as a dynamic mixture of model-based and model-free learning [3], and that the OFC may mediate this process [26]. Therefore, we considered whether our results could be explained by a change in the balance between model-based to model-free learning under OFC inactivation. This explanation has some intrinsic disadvantages, as it requires at least two parallel learning systems and a third process to integrate their outputs, i.e., it is complex, with many free parameters. We found that it was possible to reproduce our results with this approach provided we also added a forgetting parameter (Figures S3). However, the resultant fits were hard to reconcile with the general understanding of OFC function, as they did not produce a decrease in model-based learning with OFC inactivation, but rather an increase in model-free learning rates (Figure S3C). This suggests a form of structural over-fitting, consistent with the observation that the fitted parameters could not be reliably recovered from simulated data (Figure S3D). Thus, a complete or partial shift from model-based to model-free control seemed not to offer a good explanation for the experimental results.

A more promising way to account for the results is to consider the possibility that the subjects are still building, and then using, a cognitive map, but that the map is different – perhaps less precise – without the contribution of OFC. This idea would be consistent with recent arguments against pure model-free processing [27,28], evidence that the OFC is particularly important for sculpting representations of various aspects of tasks[13,29,30], and findings in OFC-lesioned macaques of impaired credit assignment [31]. Translating this idea to the current task, we hypothesized that the OFC might be particularly important for segregating and separately updating each unique cue-outcome pair, which were of uncertain importance in initial conditioning.

bioRxiv preprint doi: https://doi.org/10.1101/2022.01.25.477716; this version posted January 26, 2022. 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-NC-ND 4.0 International license.

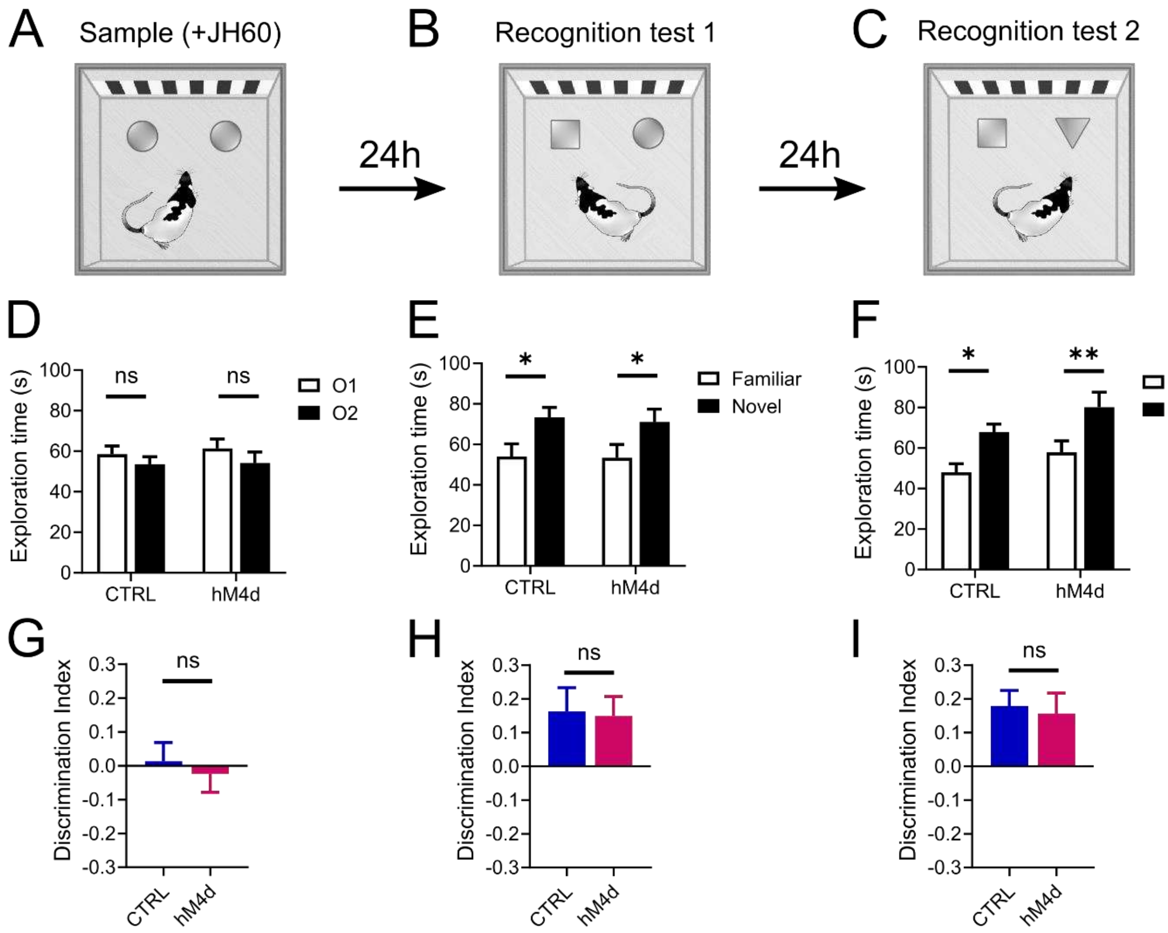

Figure 3. OFC inactivation does not affect object recognition. (A): Sample phase, where rats explored two identical objects and received JH60 injections. **(B):** First recognition test, where one familiar object was replaced by a novel one. **(C):**Second recognition test, where the previous familiar object was substituted by yet another novel object. **(D and G):** Rats in both groups explored the two objects for the same amount of time during sample (2-way ANOVA; O1 x O2: $F_{1,17} = 1.833$, $P = 0.193$; CTRL vs hM4d: $F_{1,17} = 0.14$, $P = 0.712$; interaction: $F_{1,26} = 0.059$, $P = 0.809$)(D) which was evident in the discrimination index (unpaired t-test, $P = 0.634$)(G), demonstrating that OFC inactivation does not affect exploratory behavior in this task. **(E and H):** Rats from both groups showed equally robust object recognition learning, evident in the increased exploration of the novel object (Familiar x Novel: $F_{1,17} = 13.53$, $P = 0.002$**; CTRL vs hM4d: $F_{1,17} = 0.045$, $P = 0.835$; interaction: $F_{1,26} = 0.025$, $P = 0.876$) (E) and an increase in the discrimination index, which was identical between groups ($P = 0.882$) (H), indicating that OFC inactivation in sample did not affect recognition learning or memory retention, nor did it induce some form of context-dependent learning. **(F and I):** Again, rats in both the control and hM4d groups showed a similar level of preference for the novel object (Familiar x Novel: $F_{1,17} = 18.13$, $P = 0.0005$***; CTRL vs hM4d: $F_{1,17} = 3.085$, $P = 0.097$; interaction: $F_{1,26} = 0.053$, $P = 0.82$) (F), as confirmed in the discrimination index ($P = 0.775$)(I), confirming that learning under the effects of JH60 injections was similar to when no drug was injected. Asterisks in E and F indicate results of post-hoc multiple comparison tests. Data are represented as mean ± SEM. *$P<0.05$, **$P<0.01$.

bioRxiv preprint doi: https://doi.org/10.1101/2022.01.25.477716; this version posted January 26, 2022. 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-NC-ND 4.0 International license.

We tested this proposal by fitting our data with a novel model-based reinforcement learning algorithm trained on the same sequence of trials as in the task [3,32] (Figure 4). The effect of OFC inactivation on learning during initial conditioning was captured by introducing an "imprecision" parameter (χ) that defined how credit assignment spread – i.e., whether updates were selective for each cue-outcome pair during the conditioning phase of the task (Figure 4A). Thus, receiving a banana-flavored pellet after cue A updates the association between the alternative cue B and the banana-flavored pellet by an amount proportional to χ. Only if $\chi = 0$, would the update be confined exclusively to cue A. A model with a high χ value would therefore be able to learn that auditory cues predict sucrose pellets, but would have trouble differentiating which pellet flavor (e.g., banana) is associated with which cue (A or B). Substantial confusion during conditioning (high χ) would cause the loss of value imposed by the following CTA training (Figure 4B) to be at least partially generalized to both cues A and B, due to the imprecision of specific state predictions and subsequent inference (Figure 4C), noting that the rats remained well aware of the separate values of the pellet types after the probe test (Figure 2C, right).

We found that this "imprecision" model fit our behavioral results well (Figure 5), reproducing the normal behavior in the control group and all effects of OFC inactivation, including both the apparent generalization of devaluation in the probe test (Figure 5B-C) as well as the lower asymptotic performance in conditioning (Figure 5E-F). Critical parameters in the model, particularly χ, were recoverable from simulated data (Figure S2A) [33,34]. Model fits to data from control and hM4d groups differed in their imprecision term χ, which was significantly higher in hM4d models (Figure 5B and Table S2). Furthermore, χ was highly correlated with the difference in responding to the valued (A) versus devalued (B) cues during probe (Figure 5C), even though this parameter only affected learning during conditioning (Figure 4A). Notably, this effect was not due to an effect of χ on the strength of conditioning, as these were uncorrelated (Figure 5D).

Our model also recapitulated other aspects of the results, specifically by having a value adjustment parameter (∇_{pell2cue}) that captured the asymptotic performance during conditioning. The value of this parameter differed between fits for control and hM4d subjects (Figure 5E), accounting for the reduced responding of hM4d rats at the end of conditioning (Figure 1D, 2B and 5F). Importantly, ∇_{pell2cue} did not correlate with the difference in cue responses during the probe (Figure 5G). These results confirm that the effects of OFC inactivation during model creation on subsequent model-based decision making are not related to the concurrent effects on asymptotic value estimation. The latter may be related to the known role of OFC in representing and updating outcome value [14,35].

bioRxiv preprint doi: https://doi.org/10.1101/2022.01.25.477716; this version posted January 26, 2022. 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-NC-ND 4.0 International license.

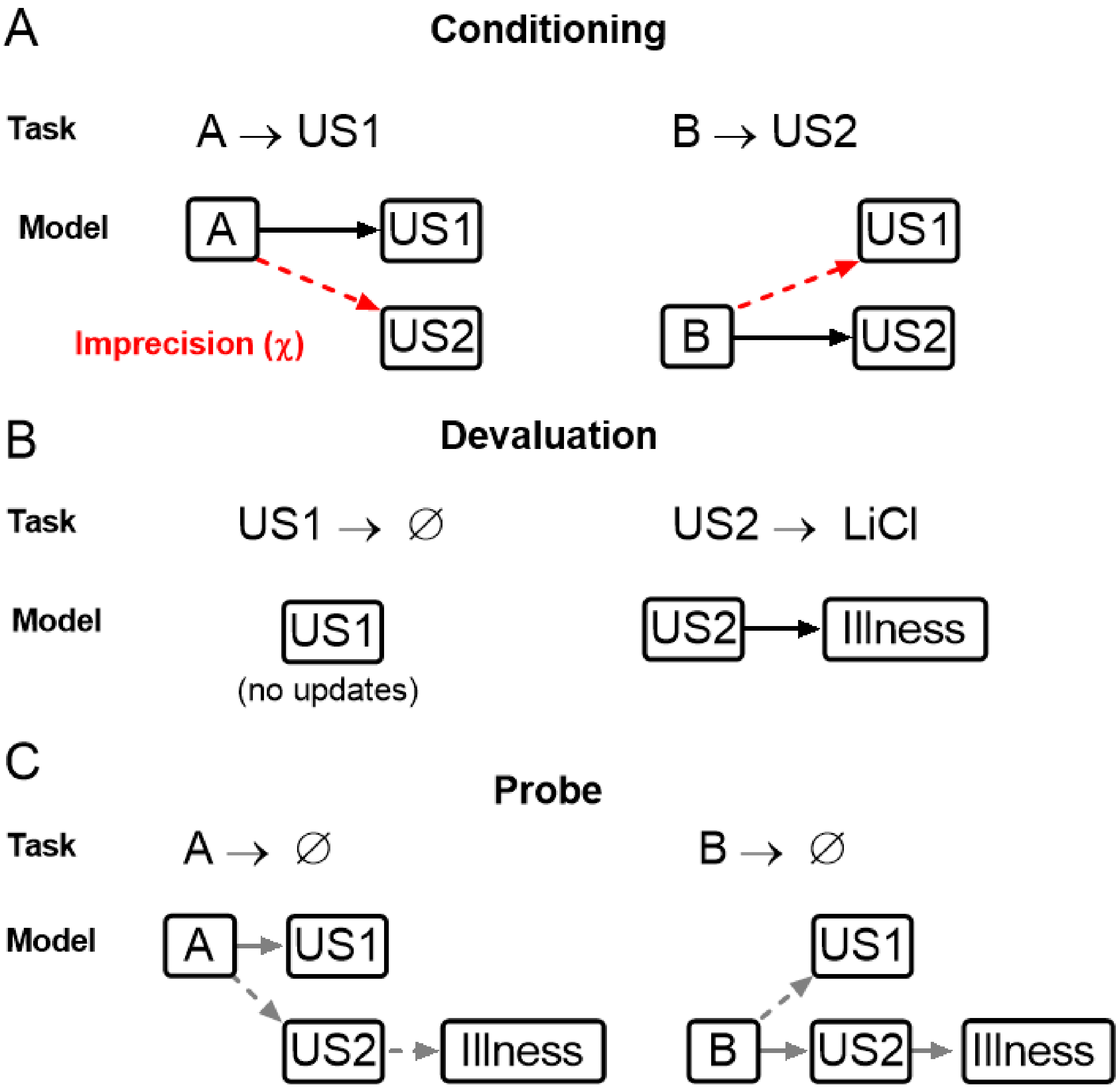

Figure 4. A model-based reinforcement learning algorithm that simulates imprecise state identity credit assignment. (A): During initial conditioning, the value and state transition matrices are updated according to the task contingencies (A-R1, B-R2; solid black arrows), except for a parallel updating of the opposite association (A-R2, B-R1), which occurs proportionately to the imprecision term χ (dashed red arrows). **(B):** During the CTA devaluation procedure, updating obeys task contingencies, with no value or state prediction updates to the R1 state, but with a learning that R2 predicts a devalued illness state. **(C)** During the probe, new learning follows the task states, and the value of cue states is adjusted according to the inferred state predictions (grey arrows), including generalized inferences driven by the imprecision term during initial acquisition (dashed grey arrows).

bioRxiv preprint doi: https://doi.org/10.1101/2022.01.25.477716; this version posted January 26, 2022. 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-NC-ND 4.0 International license.

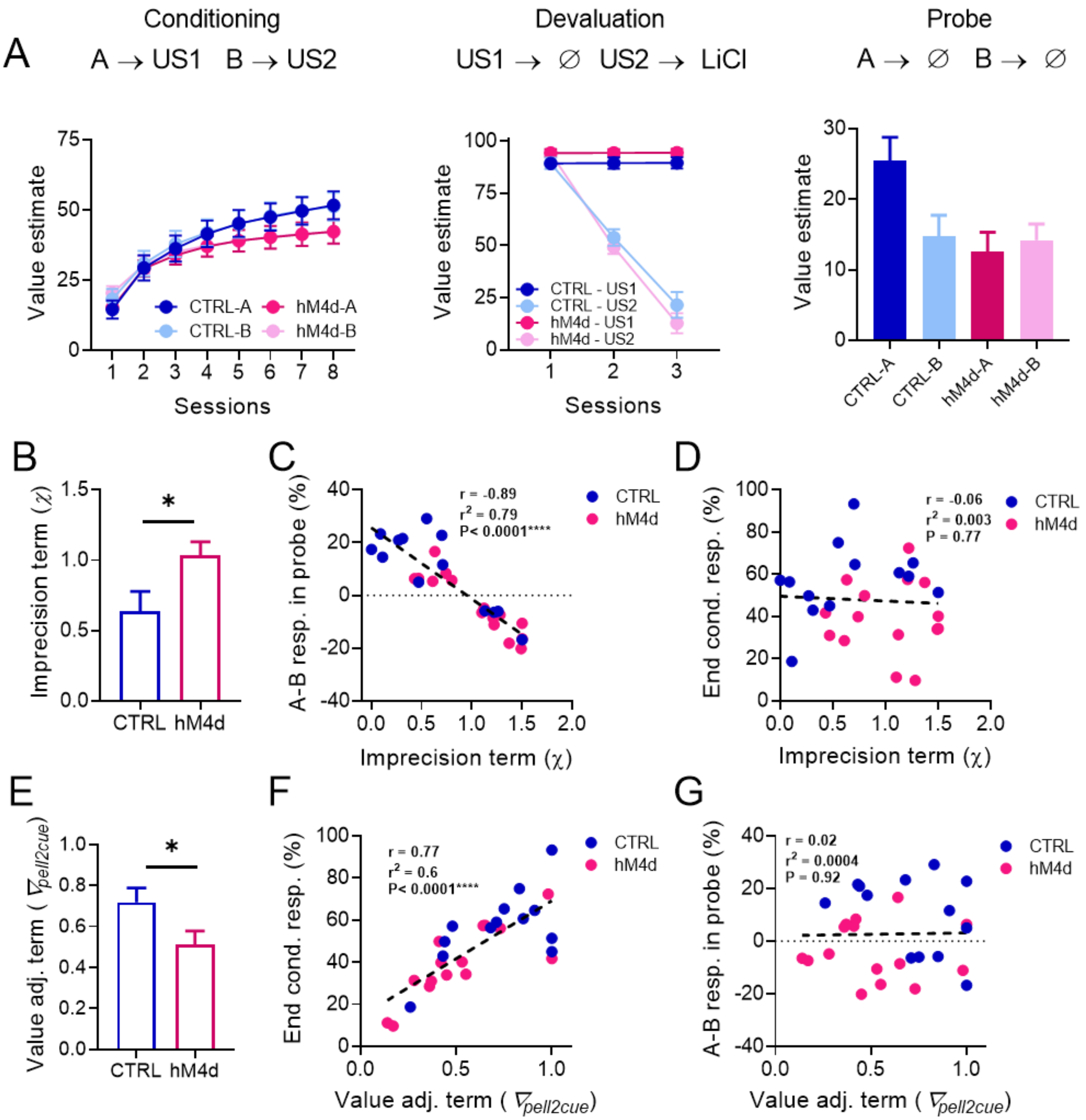

Figure 5. OFC inactivation effects on reinforcer devaluation are explained by a deficit in differentiating specific cue-outcome associations. (A): Model fit results for our model-based reinforcement learning model with potential outcome identity confusion. **(B):** The imprecision term χ was significantly higher in models fitted to hM4d behavioral data in relation to controls (P=0.027*). **(C):** χ was negatively correlated with the differential responding to cues in the probe session. **(D):** χ was not correlated with the average responding to cues at the end of conditioning. **(E):** The value adjustment term ∇_{pell2cue} was significantly lower in hM4d models (P = 0.04*). **(F):** ∇_{pell2cue} was positively correlated with average cue responding at the end of conditioning. **(G):** ∇_{pell2cue} was uncorrelated with differential responding to cues in the probe session. See Table S2 and Figure S2 for detailed parameter comparisons. Data are represented as mean ± SEM. *P<0.05.

bioRxiv preprint doi: https://doi.org/10.1101/2022.01.25.477716; this version posted January 26, 2022. 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-NC-ND 4.0 International license.

Discussion

Our study demonstrates first that OFC is necessary for the construction of a normal cognitive map and second that the OFC appears to play a circumscribed role in this construction process. In our task, the map-making apparently did not cease when OFC was inactivated, but the created map was degraded and less specific about which cues led to which outcomes. This was modeled as a lack of precision in credit assignment, but a failure to create appropriately "granular" [36,37] internal representations of these external events would produce the same result and seems more likely than a direct control of OFC over error signal assignment.

As an intuitive example of the utility of setting this granularity properly, an older child may learn that McDonald's™ serves Happy Meals™ while Burger King™ serves King Jr™ meals, each with different toys, while a younger sibling may only recall that fast food restaurants serve kids' meals. Both cognitive maps lead to food, but only one will help you collect all the Disney™ dragons! Whether to keep or discard the information related to which restaurant serves which kids' meal with which toy is a question of how to segregate the states during learning[36,37]; it is this process that we propose OFC controls or contributes to during cognitive mapping [20]. This example also illustrates the fact that the generalization afforded by discarding information is not automatically incorrect – it should respond to the exigencies of the circumstance.

We would speculate that OFC's particular contribution to this process is in determining whether to maintain separation between states that have uncertain or perhaps only potential biological significance, when other parts of the circuit might collapse them. Maps formed with too little separation due to hypofunction in OFC would tend to underrepresent potential or hidden associations and meaning and be unable to link to and infer relationships to other maps, as we have seen here. This is also evident in substance use disorder, neurodegenerative diseases, and advanced aging, in which OFC function is compromised[5,6,38–41], and in children and adolescents, which have immature frontal cortices[42,43]. Conversely, maps formed with too much separation due to an over-exhuberant OFC would tend to instill hidden meaning where it does not exist; notably such an effect is arguably evident in obsessive compulsive disorder and paranoid psychosis, which involve hyperfunction in the OFC and related areas[7,40,44–49].

The proposal that the OFC plays a critical role in defining the states that form the basis of cognitive maps is congruent with much existing data. This includes classic findings based on manipulations in the probe phase of reinforcer devaluation experiments[16–19,50], since the probe phase confounds the integration of established maps with their first time use. That is, the function proposed here would be invoked in the probe test in devaluation by the need to recognize the common reward state in the maps created during the conditioning and devaluation phases. Similar conclusions apply to other cardinal studies that have implicated the OFC in model-based behaviors, since these also normally involve integrating or changing task maps [51–53]. This more limited role for OFC also explains better why this area is necessary in many other behavioral settings where normal behavior depends upon recognizing states that are somewhat ambiguously defined with regard to biological value, including for instance the differential outcomes effect[54] , latent inhibition [22], and reversal learning[55–57], and why OFC seems to grow less important in settings like reversal learning or economic choice once maps are well-established[17,20,55,58]

bioRxiv preprint doi: https://doi.org/10.1101/2022.01.25.477716; this version posted January 26, 2022. 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-NC-ND 4.0 International license.

Finally, perhaps the most intriguing implication of our finding that OFC inactivation fails to reveal model-free learning is the possibility that most learning is, to some degree, model-based, but that mental representations or cognitive maps can be formed with different degrees of granularity or specificity, as determined by the circuits that are engaged in the learning process, including the OFC and other prefrontal areas. In the absence of experimental interventions, illness, or lesions, it could be that the main determinant of the resolution of a cognitive map would be task requirements and learning context. This would mean that perhaps there is a unified learning process that can be more or less complex depending on the contribution of specific circuits or environmental demands.

bioRxiv preprint doi: https://doi.org/10.1101/2022.01.25.477716; this version posted January 26, 2022. 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-NC-ND 4.0 International license.

Materials and Methods

Experimental Model and Subject Details

Experiments were performed on 32 male Long-Evans rats (n=16 for each group, >3 months of age at the start of the experiment, Charles River Laboratories) housed on a 12 hr light/dark cycle at 25 °C. Rats were food restricted to ~85% of their original weight for the duration of the experiments and were tested at the NIDA-IRP in accordance with NIH guidelines determined by the Animal Care and Use Committee. All rats had *ad libitum* access to water during the experiment and were fed 16-20 g of food per day, including rat chow and pellets consumed during the behavioral task. Behavior was performed during the light phase of the light/dark schedule. The number of rats used was determined based on previous publications from the lab using Pavlovian conditioning tasks. Prior to surgery, rats were handled every other day for 5-10 minutes for one week. Handling procedures included the performance of mock i.p. injections (rats were scruffed and the experimenter gently poked their abdomen with his finger or the end of a syringe with no needle attached) to prepare the subjects for future real injections. These rats were also used in another study [22]. One rat in each group was excluded due to incorrect anatomical placement, and two rats were excluded from the control group due to a hardware malfunction during one of the behavioral sessions, leading to n=13 for the control group and n=15 for the hM4d group.

Surgical procedures

Rats were anesthetized with 1-2% isoflurane and received either AAV8-CaMKIIa-hM4d-mCherry (a Gi-coupled designer receptor exclusively activated by designer drugs (DREADD)) or AAV8-hSyn-mCherry (control), both purchased from Adgene (Cambridge, MA), bilaterally into the OFC (AP −3.0 mm, ML ± 3.2 mm, and DV -4.4 and -4.5 mm from the brain surface) [22]. A total 0.5 μL was delivered in each site at 0.1 μL/min via an infusion pump.

Sensory-specific conditioning

Rats were trained and tested at least eight weeks after the surgeries in standard behavioral boxes (12” x 10” x 12,” Coulbourn Instruments, Holliston, MA). Each box was equipped with a food cup, a pellet dispenser and two wall speakers. Head entries into the food cup was measured based on breaks of an infra-red beam.

Rats were conditioned for eight sessions. Prior to each session, each rat received an i.p. injection of JH60 (0.2 mg/kg, dissolved in 0.9% NaCl) and was left in their home cage for at least 15 minutes before the start of the session, to allow for the DREADD agonist to effectively inhibit transfected OFC neurons in the hM4d group [21,22].

In every session, rats were exposed to two auditory stimuli, A and B (siren or white noise, counterbalanced across rats); each cue was presented for 10 seconds, immediately followed by the delivery of two bacon- or banana-flavored pellets (TestDiet; counterbalanced pairing). Each pairing was presented eight times per session with an average ITI of 2.5 minutes and the order of presentation was randomized and counterbalanced.

bioRxiv preprint doi: https://doi.org/10.1101/2022.01.25.477716; this version posted January 26, 2022. 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-NC-ND 4.0 International license.

Behavioral responses were quantified as the percentage of time that each rat spent in the food cup during the last 5 seconds of each CS, subtracted by the time they spent in the food cup 5 seconds before CS onset.

Reward preference tests

Prior to the devaluation procedure, rats were given a preference test comparing consumption of the two pellet-types. Rats were provided 100 pellets of each type, placed in two ceramic bowls for 30 minutes with the location of the bowls reversed every 5 minutes. The remaining pellets were counted after the 30 minute period. This procedure was repeated after the devaluation probe to confirm the permanence of conditioned taste aversion.

Reinforcer devaluation via conditioned taste aversion with LiCl

For outcome-specific reinforcer devaluation, we paired the reward associated with cue B with LiCl, while the reward associated with cue A was not paired with anything. This devaluation procedure lasted a total of six days. On days 1, 3 and 5, rats were given 30 minutes of access to the devalued pellet, followed immediately by an i.p. injection of 0.3 M LiCl, then returned to their home cages [17]. On alternate days (2, 4 and 6), rats were given 30 minutes of access to the non-devalued pellet and then returned to their home cages. All preference and consumption tests were performed in clean home cages.

Devaluation probe

The devaluation probe was performed and analyzed exactly like one of the conditioning sessions, except that no reinforcer was delivered, and the rats did not receive an injection.

Object recognition task

A subset of 10 rats from each group of the previous experiment was randomly selected for this procedure. One of the control rats was the one excluded due to incorrect anatomical placement, leading to n=9 for the control group and n=10 for the hM4d group for this experiment.

One square arena (60 x 60 cm) made of brown plexiglass with a striped black and white rectangular spatial cue was placed in a dimly (~3 lumens) red-light illuminated room. A video camera was mounted above the arenas, and activity during test sessions was digitized with a high-definition webcam (C920S PRO HD, Logitech, Suzhou, China). The objects to be discriminated were white glass bulbs, transparent glass jars, cylindrical amber glass bottles and trapezoidal white plastic bottles. All objects were glued to heavy metal disks to prevent them from being displaced by the rats, and positioned at the back corners of the arena (10 cm from walls). To avoid olfactory cues, the arena and objects were thoroughly cleaned with 0.1% acetic acid after each trial.

For habituation, the rats were positioned into the open-field arena without any objects for 10 min the day before the start of the experiment. Throughout the experiment, the position of the objects was constant, but the objects used and their relative positions were counterbalanced for every animal. In the sample phase, rats were placed in the arena facing the wall opposite the objects and were allowed to freely explore two identical objects (either two light bulbs or two jars) for 10 min. Prior to the sampling session, each rat received an i.p. injection of JH60 (0.2 mg/kg, dissolved in 0.9% NaCl) and was left in

bioRxiv preprint doi: https://doi.org/10.1101/2022.01.25.477716; this version posted January 26, 2022. 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-NC-ND 4.0 International license.

their home cage for at least 20 minutes before the start of the session. This period was given to allow for the DREADD agonist to reach the brain and effectively inhibit transfected OFC neurons in the hM4d group. After 24 h, on memory test 1, rats were allowed to explore freely one copy of the previously presented object (familiar) together with a new one (novel) for 10 min. A second memory test was performed 24 h after the first test. During the second memory test, the object that was introduced in the previous memory test was kept in place (so now it was the familiar object), and the previous familiar object was replaced by a novel object (either amber or white bottles), and rats explored freely for 10 min.

As previously described[25], exploration was defined as pointing the nose toward to an object at a distance of less than 1 cm and/or touching it with the nose. Turning around or sitting on the objects was not considered as exploratory behavior. A Discrimination Index (DI) was calculated, where DI = difference between exploration of the novel and familiar objects / total object exploration time during each memory test, such as that a DI of 0 indicates equal preference for both objects, a DI of 1 indicates exclusive exploration of the novel object, and a DI of -1 indicates exclusive exploration of the familiar object.This measure was also calculated using only the first 5 min of each test, but results were similar to when the whole test period was used (data not shown). Video recordings were scored automatically using TopScan Suite (Clever Sys, Reston, VA). Exploration times were verified manually by a trained rater blinded to treatment and objects identities using BORIS software (Version 7.9.19, University of Torino, Italy).

Histological procedures

After completion of the experiment, rats were perfused with chilled phosphate buffer saline (PBS) followed by 4% paraformaldehyde in PBS. The brains were then immersed in 18% sucrose in PBS for at least 24 hours and frozen. The brains were sliced at 40 μm and stained with DAPI (Vectashield-DAPI, Vector Lab, Burlingame, CA). Fluorescent microscopy images of the slides were acquired with a BZ-X800 Keyence microscope. Expression patterns were extracted from the images and then superimposed on anatomical templates [22].

Statistical analyses

Data were analyzed using GraphPad Prism (GraphPad Software, San Diego, CA). Error bars in figures denote the standard error of the mean. Effects of experimental variables on behavioral measures were examined with repeated-measures 2-way and 3-way ANOVAs combined with Sidak’s or Tukey’s post-hoc tests, respectively. Statistical significance threshold for all tests was set at P<0.05.

bioRxiv preprint doi: https://doi.org/10.1101/2022.01.25.477716; this version posted January 26, 2022. 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-NC-ND 4.0 International license.

Reinforcement learning modelling

Background

We modelled the five stages of the experiment in chronological order: Conditioning (COND), Preference Test 1 (PRFT1), Devaluation (DEV), Preference Test 2 (PRFT2) and finally Probe testing (PROBE). For COND and PROBE, the *Port Stay Probability (PSP)* upon cue presentation was quantified. In PRFT1, DEV and PRFT2, the *percentage of pellets eaten (PPE)* was quantified. Two pellets of a single type were delivered in each case.

On each trial, an internal value estimate ($\overline{V}$) was calculated based on contributions from a model based (MB) system (and, for the alternative hypothesis of a loss of MB learning, in combination with a model free, MF, system). This value estimate was then transformed to the behavioral measurement that was appropriate to the experimental stage. In keeping with standard practice, we described the Pavlovian connection between cue and outcome as being associations; however, in keeping with the temporal evolution of the task, we actual model them as transitions from cue to outcome. MB (and MF) systems were updated using the state transitions that were observed (e.g., A→ValuedOutcome) and the rewards that were received.

The main hypothesis (we call this Ha) that we tested was that the OFC enables precise credit assignment through separation of specific cue-outcome relations (i.e., that sound A predicts banana flavored pallets) and when deactivated, only the general relation (that any auditory cue predicts delivery of food) can be learned. However, we also tested a model (Hb) which could potentially characterize a more conventional view of OFC deactivation, namely that it would suppress MB over MF control. Since Hb mostly nests Ha, we provide an partly integrated discussion.

Formal model

$S = \{s_1, \ldots, s_n\}$ is the set of states. Each state is typically associated with the presentation of a cue or an outcome that can be rewarded or devalued, i.e., $S \sim \{A, B,$ValuedOutcome, DevaluedOutcome$\}$.

In order to be able to characterize MB and MF systems fairly, we considered forms of both that represent the uncertainty in their predictions of rewards and values. However, we adopt a heuristic Bayesian scheme, with observation rates (the equivalent of learning rates) that are parameters (rather than pure conjugate distributional updates).

Following Dearden et al. [32], normal-gamma distributions are used to characterize this uncertainty (since, following Daw et al. [3], MB and MF systems share the characterization of the values of the final, reward, states, albeit potentially with different parameters, and with only the MB system being subject to the effects of devaluation).

We write this down in terms of the value of state s. The normal-gamma distribution for the value V_s and the *precision* ρ_s^2 is written as $\mathcal{NG}(m_s, \lambda_s, \alpha_s, \beta_s)$. According to this, the *conditional* distribution of V_s given ρ_s^2 is a normal distribution

$$V_s \quad \sim \mathcal{N}(m_s, 1/(\lambda_s \rho_s^2)) \tag{1}$$

and the precision has an unconditional gamma distribution

bioRxiv preprint doi: https://doi.org/10.1101/2022.01.25.477716; this version posted January 26, 2022. 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-NC-ND 4.0 International license.

$$\rho_s^2 \quad \sim \Gamma(\alpha_s, \beta_s) \tag{2}$$

in terms of our problem, we interpret the parameters as follows:

m_s is the mean reward across the previous iterations
λ_s is the number of outcomes seen (this also includes the cases when no reward ($r = 0$) is delivered in this state)
α_s describes the total opportunity for learning about the precision; assuming that we initialize alpha to: $\alpha_s^{init} = 0.5 * \lambda_s^{init}$, then it holds that at all times $\alpha_s = 0.5 * \lambda_s$.
β_s describes the scale of the precision across previous seen rewards

implying that the marginal mean and variance of V_s are:

$$\overline{V}_s \quad = E[V_s] = m_s \qquad \text{Var}[V_s] = \frac{\beta_s}{\lambda_s*(\alpha_s-1)} \tag{3}$$

For MB computations, we also need an internal model of the state graph. We use T to describe the distribution of transition probabilities from all to all states. Programmatically, T can be described by a matrix where each row contains ϕ's that are parameters for the multinomial distribution that characterizes the transition probabilities from a "source" state s to any of the other states (including the source state itself):

$$T_{s\cdot} \sim Dirichlet(\phi_{ss_1}, \dots, \phi_{ss_n})$$

This will only be interpretable for non-terminal "source" states s, as the trial ends afterwards and no information about consecutive states can be collected. The terminal states are thus absorbing. The sum of probabilities for a fixed source state to all possible target states is 1 (see model based value calculation).

Initialization

We initialize all ϕ's in T to 10. This implies a moderately strong prior that the transition probabilities are uniform across all states:

$$\phi_{ss'}^{\text{init}} \quad = 10 \qquad \forall(s, s') \tag{4}$$

We initialize the distribution describing the reward distribution parameters to:

$$V_s^{\text{init}}, \rho_s^{2\ \text{init}} \quad \sim NG(m_s^{\text{init}} = 0, \lambda_s^{\text{init}} = 3, \alpha_s^{\text{init}} = 1.5, \beta_s^{\text{init}} = 1.5) \qquad \forall(s) \tag{5}$$

The rationale for these values is that $\alpha_s^{\text{init}} > 1$ to ensure V_s has a finite marginal variance. The value of m_s^{init} was chosen to be 0 as animals start out with no value expectation. λ_s^{init} was set to $2 \times \alpha_s^{\text{init}}$, as this ratio is also maintained by the updates. β_s^{init} was set to 1.5 in order to set the starting marginal variance to $\text{Var}\left[V_s^{\text{init}}\right] = 1$. However, we confirmed that our results are stable to quite a wide range of initialization values, provided that the variance is well-defined ($\alpha_s > 1$).

During the conditioning stage, $r_{\text{ValuedOutcome}} = r_{\text{DevaluedOutcome}} = 2$ (for the number of pellets provided). The reward of the DevaluedOutcome changes during the devaluation period to NegRew < 0, which is a parameter that captures the strength of the devaluation effect for each animal.

Model updates and value calculation

bioRxiv preprint doi: https://doi.org/10.1101/2022.01.25.477716; this version posted January 26, 2022. 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-NC-ND 4.0 International license.

The normal-gamma distribution characterizing the value V_s of a terminal state updates according to each observation. In general, given an observation $\hat{V}_s$, writing ${V'}_s, {\rho_s^2}' \sim NG({m_s}', {\lambda_s}', {\alpha_s}', {\beta_s}')$ for the updated distribution at s, we update the parameters as:

$$ {m_s}' \;\; = \frac{\lambda_s \cdot m_s + \eta \cdot \hat{V}_s}{\lambda + \eta},\ {\lambda_s}' = \lambda_s + \eta,\ {\alpha_s}' = \alpha_s + 0.5 \cdot \eta,\ {\beta_s}' = \beta_s + \frac{\eta \cdot \lambda_s \cdot (\hat{V}_s - m_s)^2}{2 * (\lambda_s + \eta)} \qquad (6) $$

where η is called an observation rate and stands in for the number of subjective observations associated with each experience – it need only be positive and is not constrained to be less than 1.

For the MB system, writing $\mathrm{V}_{s(t)}^{\mathrm{mb}} \sim \mathcal{NG}({m^{\mathrm{mb}}}_{s(t)}, {\lambda^{\mathrm{mb}}}_{s(t)}, {\alpha^{\mathrm{mb}}}_{s(t)}, {\beta^{\mathrm{mb}}}_{s(t)})$, for a terminal state, the update happens using $\hat{V}_{s(t)}^{\mathrm{mb}} = r_{s(t)}$ and observation rate $\eta = \eta^{\mathrm{mb}}$.

For the transition matrix, if the state $s(t)$ is a non-terminal state that is followed by state $s(t+1)$, the parameters of the transition probability distribution $T_{s(t)\cdot}$ are updated using a notional transition observation rate η^{t} as:

$$ {\phi'}_{s(t)s(t+1)} = \phi_{s(t)s(t+1)} + \eta^{\mathrm{t}} \qquad (7) $$

The MB system combines its knowledge of transitions and immediate rewards by applying the Bellman equation, which, in this case is very straightforward, since there are only two steps. Ignoring any posterior correlation between T and μ, σ, this implies that:

$$ \overline{V}_{s(t)}^{\mathrm{mb}} = \begin{cases} m_{s(t)}^{\mathrm{mb}} & \text{if } s(t) \text{ is a terminal state} \\ m_{s(t)}^{\mathrm{mb}} + \gamma^{\mathrm{mb}} \cdot \sum_{s(t+1)} E\left[T_{s(t)s(t+1)}\right] \cdot m_{s(t+1)}^{\mathrm{mb}} & \text{otherwise} \end{cases} $$

The expected value for the next state is discounted by γ^{mb}, which normally is close to 1. The expected value for the transition probability from state $s(t)$ to state $s(t+1)$ can be calculated using: $E[T_{s(t)s(t+1)}] = \phi_{s(t)s(t+1)} / \sum_\omega \phi_{s(t)\omega}$.

The approximate variance can be calculated from the Bellman equation (again ignoring correlations).

Transformation of Estimated Values to Behavioral Measures

Having generated a prediction $\overline{V}_{s(t)}^{\mathrm{mb}}$ from the MB system, it is necessary to convert it into the different experimental measures used in the various stages of the experimental paradigm. To do this, the combined value is normalized by the standard scalar reward received (2, for the number of pellets), and thresholded at 0 in order to avoid negative percentages when calculating the behavioral measures:

$$ \overline{V}_{s(t)}^{\mathrm{norm}} \;\; = \max\left(\frac{\overline{V}_{s(t)}^{\mathrm{mb}}}{2}, 0\right) \qquad (8) $$

This normalized value can then be transformed to the respective behavioral measures for each stage, each given as percentages in the range [0,100]:

bioRxiv preprint doi: https://doi.org/10.1101/2022.01.25.477716; this version posted January 26, 2022. 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-NC-ND 4.0 International license.

$$\mathrm{PSP}^{\mathrm{COND}}_{s(t)} = \overline{V}^{\mathrm{norm}}_{s(t)} \cdot 100 \cdot \nabla_{\mathrm{pell2cue}} \tag{9}$$

$$\mathrm{PPE}^{\mathrm{DEV}}_{s(t)} = \overline{V}^{\mathrm{norm}}_{s(t)} \cdot 100 \tag{10}$$

$$\mathrm{PSP}^{\mathrm{PROBE}}_{s(t)} = \overline{V}^{\mathrm{norm}}_{s(t)} \cdot 100 \cdot \nabla_{\mathrm{pell2cue}} \cdot \nabla_{\mathrm{cp}} \tag{11}$$

$\nabla_{\mathrm{pell2cue}}$ accounts for the difference in the impact of a secondary predictor versus a primary reinforcer, and ∇_{cp} may account for the forgetting of cue values from COND to the PROBE phase. Both factors are in the range [0,1]. An additional factor for the calculation of $\mathrm{PPE}^{\mathrm{DEV}}_{s}$ was not necessary. $\mathrm{PPE}^{\mathrm{PRFT1}}_{s}$ and $\mathrm{PPE}^{\mathrm{PRFT2}}_{s}$ are calculated the same way as $\mathrm{PPE}^{\mathrm{DEV}}_{s}$.

Ha: Outcome-specific encoding deficit

In this version, only the MB system is used, and we assume no forgetting happens from COND to PROBE so ∇_{cp} is fixed to 1.

We model the inactivation of OFC as implying that the representation of the relevant cues (here, A and B) is potentially only partially distinct. Thus, if, for instance $s(t) = A$ is presented, then writing $\tilde{s}(t) = B$ as the 'other' cue, we imagine a spillover or fuzziness factor χ is introduced that is taken into consideration when doing the updates so that, along with equation 7, we have

$$\phi'_{\tilde{s}(t)s(t+1)} = \phi_{\tilde{s}(t)s(t+1)} + \eta^{\mathrm{t}}\chi \tag{12}$$

If $\chi = 0$, nothing is learned for the opposite state, if $\chi = 1$, then exactly the same transition information is learned for both states, and if $\chi > 1$, then more is learned for the opposite/unseen state. Note that we continue to consider the outcome pellets to be perfectly distinguishable.

The free parameters used for model fitting are: NegRew, $\nabla_{\mathrm{pell2cue}}$, η^{mb}, η^{t}, χ.

Model Fitting

Separate sets of parameters were fit for each animal using scipy.optimize.least_squares, optimizing the mean squared error (MSE) between the real behavioral recordings and the model "behavior" outputs based on the current set of parameters. A weighted MSE was used in order to increase the contribution of the PROBE trials as behavioral differences across groups (control/OFC deactivation) were most apparent here, and the number of trials comparably few (there is 8x more condition trials, so PROBE trials have an 8x higher weight). The following bounds for the parameter fitting were defined as follows:

Param	NegRew	$\nabla_{\mathrm{pell2cue}}$	η^{mb}	η^{t}	χ
Min	-90	0	0	0	0
Max	0	1	40	40	1.5

bioRxiv preprint doi: https://doi.org/10.1101/2022.01.25.477716; this version posted January 26, 2022. 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-NC-ND 4.0 International license.

Individual parameter estimates for either of the models were then compared across groups using t-tests and Bonferroni-corrected for multiple comparisons.

Parameter Recovery

In order to ensure that recovered parameter values are meaningful in case of the model fits, we checked parameter recoverability. Here, we use known parameter values along with realistic noise to generate synthetic data, and then assess if we can recover from these data values of the parameters that are close to the original generating levels. In order to stay close to the real data, we used the parameters recovered for each animal individually to generate one synthetic dataset/behavioral trace per animal. The noise was generated using individual variability estimates of per trial behavioral measures for each experiment stage (COND, DEV, PROBE). This yields 28 pairs (one pair per animal) of real and estimated parameter values for each of the model's parameters. Good parameter recoverability is when real and estimated parameter values are well correlated.

Recovery of most of the parameters was good ($r_{\text{NegRew}} = 0.9$, $r_{\nabla_{\text{pell2cue}}} = 0.8$, $r_{\eta^{\text{mb}}} = 0.8$, and $r_{\chi} = 0.7$); only the recovery of the state transition observation rate η^{t} was slightly less faithful ($r_{\eta^{\text{t}}} = 0.6$), and so should be interpreted cautiously.

Repeating the recovery procedure multiple times produced comparable results. We also used a synthetic generative procedure to assess the posterior correlations between recovered parameter values, something that matters for prediction, albeit less for the overall interpretation of the model. We started out with the median parameter values across animals to generate synthetic data, with noise generated based on the variability of behavioral measures per experiment stage, this time on the group level, and recovered those parameter values from these data. We did this 30 times and assessed the correlations between all pairs of inferred parameters. We found that most of the correlations were mild – although the highest correlations between ∇_{pell2cue} and η^{t} ($r = -0.57$), were quite substantial. This is not unexpected, as in effect ∇_{pell2cue} accounts for the difference between the asymptotic performance at the end of conditioning, which is in turn set by the observation rates.

Hypothesis Hb. MB deficit

Hb parameterizes a more conventional view of the effect of OFC inactivation, allowing for a combination between MF and MB learning and control, with the possibility that this combination is disturbed by inactivation.

As hypothesis Hb makes use of both model free and model based value systems, it employs two sets of value distributions: $V_s^{\text{mf}}, \rho_s^{2\ \text{mf}}$ and $V_s^{\text{mb}}, \rho_s^{2\ \text{mb}}$. MB learning and inference happens as for hypothesis Ha, except that the imprecision parameter χ is not part of Hb. Following Dearden et al. (18), the MF value system uses normal-gamma distributions for characterizing the values V_s^{mf} of all states s, both terminal (with rewards) and non-terminal (with cues).

For the MF system, each time the animal passes through state s, the value distribution at this state is updated according to either a scalar estimate $\hat{V}_s$ of the long-run reward from that state s for the MF system, or the immediate reward r using an observation rate η^{mf}.

bioRxiv preprint doi: https://doi.org/10.1101/2022.01.25.477716; this version posted January 26, 2022. 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-NC-ND 4.0 International license.

Updating the MF values of terminal states is the same as for the MB system (using equation 6) with $\hat{V}^{\mathrm{mf}}_{s(t)} = r_{s(t)}$ and an observation rate $\eta = \eta^{\mathrm{mf}}$. Updating the values of non-terminal (cue) states also follows equation (6), but now (since, in this task, there are no rewards at non-terminal states) with

$$\hat{V}^{\mathrm{mf}}_{s(t)} = \gamma^{\mathrm{mf}} \cdot \overline{V}^{\mathrm{mf}}_{s(t+1)}$$

Generally, the estimated value of the model free system is $\overline{V}^{\mathrm{mf}}_{s} = m^{\mathrm{mf}}_{s}$ and the estimated variance is given by the expression in equation 3.

According to Hb, both MB and MF contribute to the value of a cue, according to a convex combination parameter w^{mf}, which is in range [0,1] with 0 meaning only the model-based system is used and 1 that only the model free system is used:

$$\overline{V}^{\mathrm{comb}}_{s(t)} = w^{\mathrm{mf}} \cdot \overline{V}^{\mathrm{mf}}_{s(t)} + (1 - w^{\mathrm{mf}}) \cdot \overline{V}^{\mathrm{mb}}_{s(t)}, \tag{13}$$

This then generates the normalized value

$$\overline{V}^{\mathrm{norm}}_{s(t)} = \max\left(\frac{\overline{V}^{\mathrm{comb}}_{s(t)}}{2}, 0\right) \tag{14}$$

which leads to behavioral measures as in equations (9)-(11).

For convenience of fitting, the observation rate for the transition matrix was fixed to the one for the model based value distributions $\eta^{\mathrm{t}} = \eta^{\mathrm{mb}}$, and γ^{mf} and γ^{mb} were set to 1. The free parameters used for model fitting were therefore: NegRew, $\overline{V}_{\mathrm{pell2cue}}$, $\overline{V}_{\mathrm{cp}}$, η^{mf}, η^{mb} and w^{mf}. As an important simplification, we fixed w^{mf} to have the same value for COND and PROBE, even in the inactivation case, as if this had been stamped in during COND, for instance because of heightened MB uncertainty. If w^{mf} was lower in PROBE then, we would not have expected such equivalent decreased responding to both cues. An alternative possibility we did not explore is that inactivation would leave the MB system with impaired learning in COND, even at asymptote for both cues; and that if w^{mf} was indeed lower in PROBE, reduced responding would come from averaging a persistent value from the MF system with the decreased output of the MB system. This would be an alternative to making parameter $\overline{V}_{\mathrm{cp}}$ small.

The same constraints as above were used for fitting the MB system (albeit with χ effectively clamped at 0). Additionally, we had

Param	$\overline{V}_{\mathrm{cp}}$	η^{mf}	w^{mf}
Min	0	0	0
Max	1	40	1

bioRxiv preprint doi: https://doi.org/10.1101/2022.01.25.477716; this version posted January 26, 2022. 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-NC-ND 4.0 International license.

Parameter recovery of the observation rate parameters were the least faithful ($r_{\eta^{\mathrm{mf}}} = 0.6$, $r_{\eta^{\mathrm{mb}}} = 0.2$ and $r_{w^{\mathrm{mf}}} = 06$), while the estimated values of the other parameters were closer to real ones ($r_{\mathrm{NegRew}} = 0.9$, $r_{\nabla_{\mathrm{pell2cue}}} = 0.9$, $r_{\nabla_{\mathrm{cp}}} = 0.8$). Thus, when interpreting this model, less emphasis should be placed on the first three parameters. Correlations in the recovered values of the parameters were mild – with the highest correlation being between $\nabla_{\mathrm{pell2cue}}$ and w^{mf} ($r = -0.54$).

bioRxiv preprint doi: https://doi.org/10.1101/2022.01.25.477716; this version posted January 26, 2022. 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-NC-ND 4.0 International license.

References

1. Tolman, E. C. Cognitive maps in rats and men. *Psychol. Rev.* **55**, 189–208 (1948).
2. Behrens, T. E. J. *et al.* What Is a Cognitive Map? Organizing Knowledge for Flexible Behavior. *Neuron* **100**, 490–509 (2018).
3. Daw, N. D., Niv, Y. & Dayan, P. Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioral control. *Nat. Neurosci. 2005 812* **8**, 1704–1711 (2005).
4. Titone, D., Ditman, T., Holzman, P. S., Eichenbaum, H. & Levy, D. L. Transitive inference in schizophrenia: impairments in relational memory organization. *Schizophr. Res.* **68**, 235–247 (2004).
5. Schoenbaum, G., Chang, C. Y., Lucantonio, F. & Takahashi, Y. K. Thinking Outside the Box: Orbitofrontal Cortex, Imagination, and How We Can Treat Addiction. *Neuropsychopharmacology* vol. 41 2966–2976 (2016).
6. Shields, C. N. & Gremel, C. M. Review of Orbitofrontal Cortex in Alcohol Dependence: A Disrupted Cognitive Map? *Alcohol. Clin. Exp. Res.* **44**, 1952–1964 (2020).
7. Sharp, P. B., Dolan, R. J. & Eldar, E. Disrupted state transition learning as a computational marker of compulsivity. (2020) doi:10.31234/OSF.IO/X29JQ.
8. Stalnaker, T. A., Cooch, N. K. & Schoenbaum, G. What the orbitofrontal cortex does not do. *Nature Neuroscience* vol. 18 620–627 (2015).
9. Wallis, J. D. Cross-species studies of orbitofrontal cortex and value-based decision-making. *Nat. Neurosci. 2011 151* **15**, 13–19 (2011).
10. Rudebeck, P. H. & Rich, E. L. Orbitofrontal cortex. *Curr. Biol.* **28**, R1083–R1088 (2018).
11. 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).
12. Wilson, R. C., Takahashi, Y. K., Schoenbaum, G. & Niv, Y. Orbitofrontal cortex as a cognitive map of task space. *Neuron* **81**, 267–279 (2014).
13. Schuck, N. W., Cai, M. B., Wilson, R. C. & Niv, Y. Human Orbitofrontal Cortex Represents a Cognitive Map of State Space. *Neuron* **91**, 1402–1412 (2016).
14. Rustichini, A. & Padoa-Schioppa, C. A neuro-computational model of economic decisions. *J. Neurophysiol.* **114**, 1382–1398 (2015).
15. Howard, J. D. & Kahnt, T. To be specific: The role of orbitofrontal cortex in signaling reward identity. *Behav. Neurosci.* **135**, 210–217 (2021).
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. 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).
18. Izquierdo, A., Suda, R. K. & Murray, E. A. Bilateral Orbital Prefrontal Cortex Lesions in Rhesus Monkeys Disrupt Choices Guided by Both Reward Value and Reward Contingency. *J. Neurosci.* **24**, 7540–7548 (2004).

bioRxiv preprint doi: https://doi.org/10.1101/2022.01.25.477716; this version posted January 26, 2022. 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-NC-ND 4.0 International license.

19. Howard, J. D. *et al.* Targeted Stimulation of Human Orbitofrontal Networks Disrupts Outcome-Guided Behavior. *Curr. Biol.* **30**, 490-498.e4 (2020).

20. Gardner, M. P. H. & Schoenbaum, G. The orbitofrontal cartographer. *Behav. Neurosci.* **135**, 267–276 (2021).

21. Bonaventura, J. *et al.* High-potency ligands for DREADD imaging and activation in rodents and monkeys. *Nat. Commun.* **10**, 1–12 (2019).

22. Costa, K. M., Sengupta, A. & Schoenbaum, G. The orbitofrontal cortex is necessary for learning to ignore. *Curr. Biol.* **31**, 2652-2657.e3 (2021).

23. Gomez, J. L. *et al.* Chemogenetics revealed: DREADD occupancy and activation via converted clozapine. *Science (80-.).* **357**, 503–507 (2017).

24. Panayi, M. C. & Killcross, S. The Role of the Rodent Lateral Orbitofrontal Cortex in Simple Pavlovian Cue-Outcome Learning Depends on Training Experience. **2**, 1–14 (2021).

25. Weiler, M. *et al.* Effects of repetitive Transcranial Magnetic Stimulation in aged rats depend on pre-treatment cognitive status: Toward individualized intervention for successful cognitive aging. *Brain Stimul. Basic, Transl. Clin. Res. Neuromodulation* **14**, 1219–1225 (2021).

26. Panayi, M. C., Khamassi, M. & Killcross, S. The rodent lateral orbitofrontal cortex as an arbitrator selecting between model-based and model-free learning systems. *Behav. Neurosci.* **135**, 226–244 (2021).

27. Dezfouli, A. & Balleine, B. W. Habits, action sequences and reinforcement learning. *Eur. J. Neurosci.* **35**, 1036–1051 (2012).

28. Miller, K. J., Shenhav, A. & Ludvig, E. A. Habits without values. *Psychol. Rev.* **126**, 292–311 (2019).

29. Zhou, J. *et al.* Evolving schema representations in orbitofrontal ensembles during learning. *Nat. 2020 5907847* **590**, 606–611 (2020).

30. Takahashi, Y. K., Stalnaker, T. A., Marrero-Garcia, Y., Rada, R. M. & Schoenbaum, G. Expectancy-Related Changes in Dopaminergic Error Signals Are Impaired by Cocaine Self-Administration. *Neuron* **101**, 294-306.e3 (2019).

31. 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).

32. Dearden, R., Dearden, R., Friedman, N. & Russell, S. Bayesian Q-learning. *IN AAAI/IAAI* 761--768 (1998).

33. Wilson, R. C. & Collins, A. G. E. Ten simple rules for the computational modeling of behavioral data. *Elife* **8**, (2019).

34. Palminteri, S., Wyart, V. & Koechlin, E. The Importance of Falsification in Computational Cognitive Modeling. *Trends Cogn. Sci.* **21**, 425–433 (2017).

35. Camille, N., Griffiths, C. A., Vo, K., Fellows, L. K. & Kable, J. W. Ventromedial Frontal Lobe Damage Disrupts Value Maximization in Humans. *J. Neurosci.* **31**, 7527–7532 (2011).

36. Dezfouli, A. & Balleine, B. W. Learning the structure of the world: The adaptive nature of state-space and action representations in multi-stage decision-making. *PLOS Comput.*

bioRxiv preprint doi: https://doi.org/10.1101/2022.01.25.477716; this version posted January 26, 2022. 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-NC-ND 4.0 International license.

Biol. **15**, e1007334 (2019).

37. Gershman, S. J. & Niv, Y. Learning latent structure: carving nature at its joints. *Curr. Opin. Neurobiol.* **20**, 251–256 (2010).

38. Denburg, N. L. *et al.* The Orbitofrontal Cortex, Real-World Decision Making, and Normal Aging. *Ann. N. Y. Acad. Sci.* **1121**, 480 (2007).

39. Van Hoesen, G. W., Parvizi, J. & Chu, C. C. Orbitofrontal Cortex Pathology in Alzheimer's Disease. *Cereb. Cortex* **10**, 243–251 (2000).

40. Jackowski, A. P. *et al.* The involvement of the orbitofrontal cortex in psychiatric disorders: an update of neuroimaging findings. *Brazilian J. Psychiatry* **34**, 207–212 (2012).

41. Volkow, N. D. & Fowler, J. S. Addiction, a Disease of Compulsion and Drive: Involvement of the Orbitofrontal Cortex. *Cereb. Cortex* **10**, 318–325 (2000).

42. Decker, J. H., Otto, A. R., Daw, N. D. & Hartley, C. A. From Creatures of Habit to Goal-Directed Learners: Tracking the Developmental Emergence of Model-Based Reinforcement Learning. *Psychol. Sci.* **27**, 848–858 (2016).

43. Sowell, E. R., Thompson, P. M., Holmes, C. J., Jernigan, T. L. & Toga, A. W. In vivo evidence for post-adolescent brain maturation in frontal and striatal regions. *Nat. Neurosci. 1999 210* **2**, 859–861 (1999).

44. Fradkin, I., Adams, R. A., Parr, T., Roiser, J. P. & Huppert, J. D. Searching for an anchor in an unpredictable world: A computational model of obsessive compulsive disorder. *Psychol. Rev.* **127**, (2020).

45. Jung, W. H. *et al.* Abnormal corticostriatal-limbic functional connectivity in obsessive–compulsive disorder during reward processing and resting-state. *NeuroImage Clin.* **3**, 27–38 (2013).

46. Rauch, S. L. *et al.* Functional Magnetic Resonance Imaging Study of Regional Brain Activation During Implicit Sequence Learning in Obsessive–Compulsive Disorder. *Biol. Psychiatry* **61**, 330–336 (2007).

47. Walther, S. *et al.* Limbic links to paranoia: increased resting-state functional connectivity between amygdala, hippocampus and orbitofrontal cortex in schizophrenia patients with paranoia. *Eur. Arch. Psychiatry Clin. Neurosci.* **1**, 1–12 (2021).

48. Yu, L., Kazinka, R., Pratt, D., Kwashie, A. & Macdonald, A. W. Resting-State Networks Associated with Behavioral and Self-Reported Measures of Persecutory Ideation in Psychosis. *Brain Sci. 2021, Vol. 11, Page 1490* **11**, 1490 (2021).

49. Pauly, K. *et al.* Cerebral Dysfunctions of Emotion—Cognition Interactions in Adolescent-Onset Schizophrenia. *J. Am. Acad. Child Adolesc. Psychiatry* **47**, 1299–1310 (2008).

50. 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).

51. Jones, J. L. *et al.* Orbitofrontal cortex supports behavior and learning using inferred but not cached values. *Science (80-.).* **338**, 953–956 (2012).

52. Takahashi, Y. K. *et al.* The Orbitofrontal Cortex and Ventral Tegmental Area Are Necessary for Learning from Unexpected Outcomes. *Neuron* **62**, 269–280 (2009).

53. Ostlund, S. B. & Balleine, B. W. Orbitofrontal Cortex Mediates Outcome Encoding in Pavlovian But Not Instrumental Conditioning. *J. Neurosci.* **27**, 4819–4825 (2007).

bioRxiv preprint doi: https://doi.org/10.1101/2022.01.25.477716; this version posted January 26, 2022. 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-NC-ND 4.0 International license.

54. McDannald, M. A., Saddoris, M. P., Gallagher, M. & Holland, P. C. Lesions of Orbitofrontal Cortex Impair Rats' Differential Outcome Expectancy Learning But Not Conditioned Stimulus-Potentiated Feeding. *J. Neurosci.* **25**, 4626–4632 (2005).

55. Schoenbaum, G., Nugent, S. L., Saddoris, M. P. & Setlow, B. Orbitofrontal lesions in rats impair reversal but not acquisition of go, no-go odor discriminations. *Neuroreport* **13**, 885–890 (2002).

56. Jones, B. & Mishkin, M. Limbic lesions and the problem of stimulus—Reinforcement associations. *Exp. Neurol.* **36**, 362–377 (1972).

57. Ghods-Sharifi, S., Haluk, D. M. & Floresco, S. B. Differential effects of inactivation of the orbitofrontal cortex on strategy set-shifting and reversal learning. *Neurobiol. Learn. Mem.* **89**, 567–573 (2008).

58. Dias, R., Robbins, T. W. & Roberts, A. C. Dissociable Forms of Inhibitory Control within Prefrontal Cortex with an Analog of the Wisconsin Card Sort Test: Restriction to Novel Situations and Independence from "On-Line" Processing. *J. Neurosci.* **17**, 9285–9297 (1997).

bioRxiv preprint doi: https://doi.org/10.1101/2022.01.25.477716; this version posted January 26, 2022. 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-NC-ND 4.0 International license.

Acknowledgments: We thank Jordi Bonaventura for guidance on chemogenetic methods, Nishika Raheja and Savana Agyemang for technical assistance, as well as Evan Hart and Marios Panayi for stimulating discussions. The opinions expressed in this article are the authors' own and do not reflect the view of the NIH/DHHS.

Funding:

Intramural Research Program at the National Institute on Drug Abuse (KMC, MG, GS)

Max Planck Society (RS, KL, PD)

Humboldt Foundation (PD).

German Research Foundation grant MA 8509/1-1 (KMC)

Author contributions:

Conceptualization: KMC, MPHG, PD, GS

Methodology: KMC, RS, KL, MPHG, PD, GS

Investigation: KMC

Software: RS, KL, PD

Validation: KMC, RS, KL, PD

Data Curation: KMC, RS

Formal analysis: KMC, RS

Visualization: KMC, RS

Resources: PD, GS

Supervision: PD, GS

Project Management: PD, GS

Writing – original draft: KMC

Writing – review & editing: KMC, RS, PD, GS

Competing interests: Authors declare that they have no competing interests.

Data and materials availability: all code and data used in this study are available on https://colab.research.google.com/drive/1VYRAnvAO8OmzQpVaJe5radKIZnpEn638?usp=sharing. Additional information on materials and protocols are available upon request to the corresponding authors.

bioRxiv preprint doi: https://doi.org/10.1101/2022.01.25.477716; this version posted January 26, 2022. 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-NC-ND 4.0 International license.

Supplementary Information

Table S1. Statistical results of behavioral experiments

Conditioning (3-way ANOVA)	***SS***	***MS***	***F (DFn, DFd)***	***P value***
Sessions	39952	5707	F (7, 182) = 26.74	P<0.0001****
(CTRL vs hM4d)	2512	2512	F (1, 26) = 0.7170	P=0.4048
(A vs B)	12.16	12.16	F (1, 26) = 0.02112	P=0.8856
Sessions x (CTRL vs hM4d)	6981	997.3	F (7, 182) = 4.672	P<0.0001****
Sessions x (A vs B)	874.0	124.9	F (7, 182) = 1.353	P=0.2279
(CTRL vs hM4d) x (A vs B)	9.810	9.810	F (1, 26) = 0.01703	P=0.8972
Sessions x (CTRL vs hM4d) x (A vs B)	517.3	73.90	F (7, 182) = 0.8008	P=0.5876

CTA (3-way ANOVA)	***SS***	***MS***	***F (DFn, DFd)***	***P value***
Sessions	37224	18612	F (2, 52) = 64.18	P<0.0001****
(CTRL vs hM4d)	1344	1344	F (1, 26) = 2.912	P=0.0998
(ND vs D)	53832	53832	F (1, 26) = 127.1	P<0.0001****
Sessions x (CTRL vs hM4d)	39.38	19.69	F (2, 52) = 0.06789	P=0.9344
Sessions x (ND vs D)	41521	20761	F (2, 52) = 83.36	P<0.0001****
(CTRL vs hM4d) x (ND vs D)	155.6	155.6	F (1, 26) = 0.3675	P=0.5497
Sessions x (CTRL vs hM4d) x (ND vs D)	32.92	16.46	F (2, 52) = 0.06609	P=0.9361

PROBE (2-way ANOVA)	***SS***	***MS***	***F (DFn, DFd)***	***P value***
(CTRL vs hM4d)	806.6	806.6	F (1, 26) = 4.340	P=0.0472*
(A vs B)	143.3	143.3	F (1, 26) = 1.743	P=0.1983
(CTRL vs hM4d) x (A vs B)	658.8	658.8	F (1, 26) = 8.013	P=0.0088**

bioRxiv preprint doi: https://doi.org/10.1101/2022.01.25.477716; this version posted January 26, 2022. 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-NC-ND 4.0 International license.

Table S2. Comparisons of fitted parameters for the control and hM4d groups withing the two tested reinforcement learning models. Data are represented as mean ± SEM.

Ha: Precision deficit model parameters	**Control**	**hM4d**	**P value**
Imprecision term - χ	0.64 ± 0.138	1.031 ± 0.099	P=0.027*
Model-based value observation rate - η^{mb}	2.669 ± 1.964	5.062 ± 2.626	P=0.169
Model-based transition observation rate - η^{tm}	11 ± 4.657	22.41 ± 4.477	P=0.067
Strength of devaluation - NegRew	-50.82 ± 5.092	-60.21 ± 4.81	P=0.185
Value adjustment - $\nabla_{pell2cue}$	0.718 ± 0.069	0.512 ± 0.066	P=0.04*

Hb: MB vs MF model parameters	***Control***	***hM4d***	***P value***
Model-free observation rate - η^{mf}	3.71 ± 2.778	15.21 ± 4.758	P=0.007**
Model-based observation rate - η^{mb}	12.8 ± 4.51	15.77 ± 4.82	P=0.575
Contribution of model-free system - w^{mf}	0.553 ± 0.106	0.703 ± 0.081	P=0.513
Strength of devaluation - NegRew	-52.09 ± 5.85	-54.78 ± 5.243	P=0.821
Value adjustment - $\nabla_{pell2cue}$	0.761 ± 0.069	0.501 ± 0.059	P=0.012*
Conditioning to probe forgetting - ∇_{cp}	0.603 ± 0.079	0.385 ± 0.083	P=0.071

bioRxiv preprint doi: https://doi.org/10.1101/2022.01.25.477716; this version posted January 26, 2022. 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-NC-ND 4.0 International license.

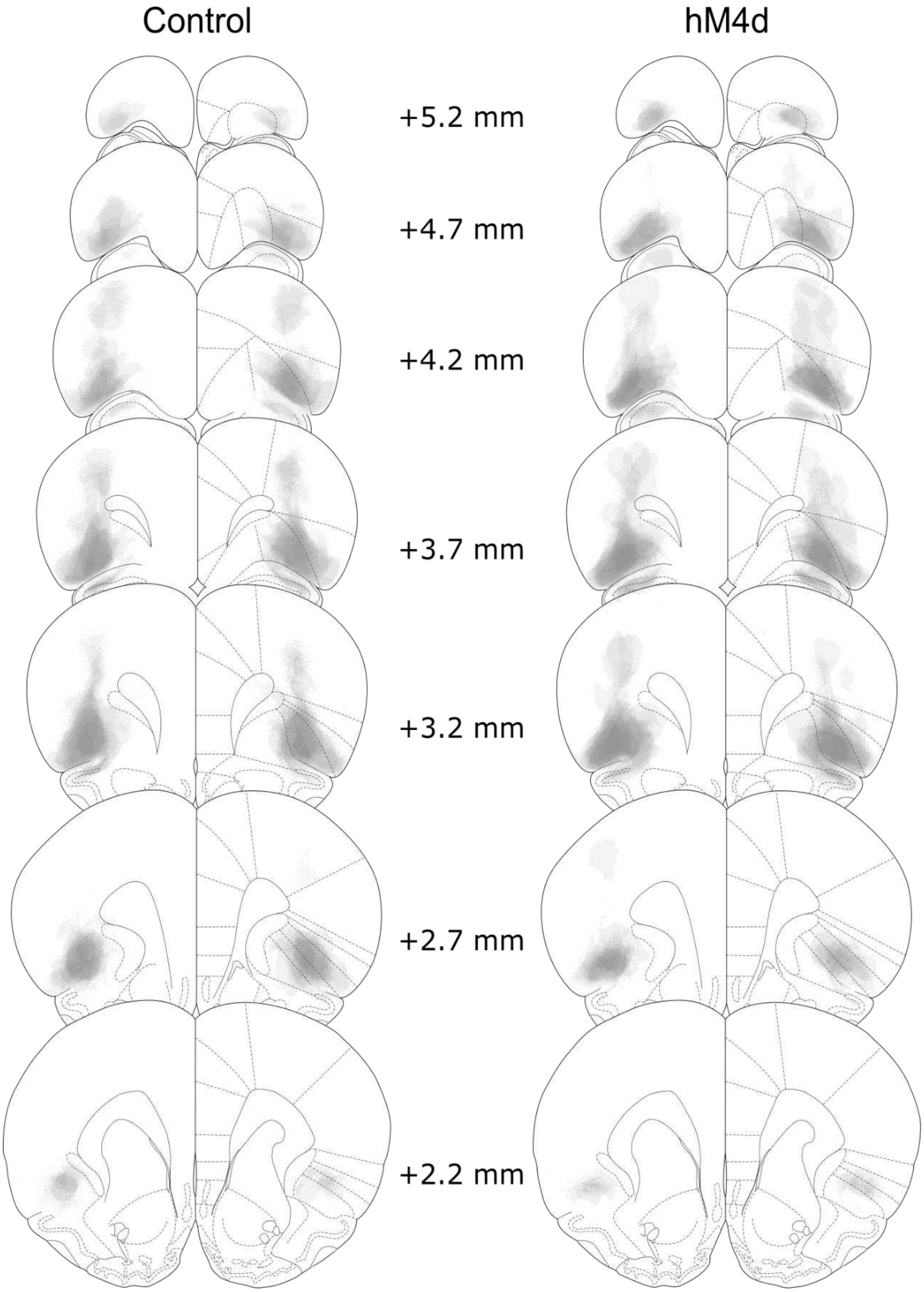

Figure S1. Histological validation of DREADD strategy. Reconstruction of viral expression patterns in the OFC across the control and hM4d groups. Viral spread was mostly contained withing OFC and was similar for control and hM4d subjects.

bioRxiv preprint doi: https://doi.org/10.1101/2022.01.25.477716; this version posted January 26, 2022. 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-NC-ND 4.0 International license.

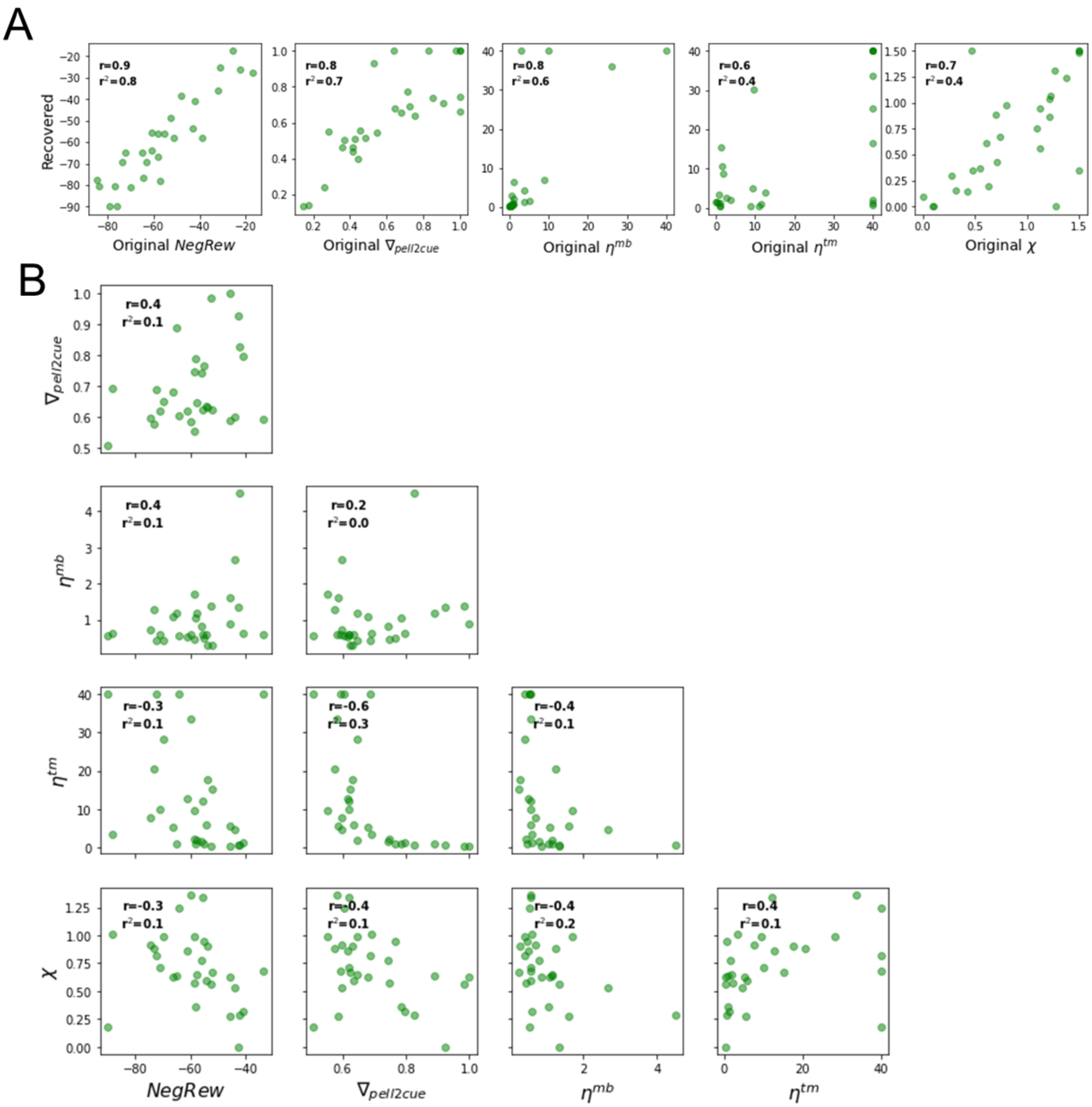

Figure S2. Parameter recovery and correlations for the reinforcement learning model with association specificity deficit. A: Correlations between estimated and original parameters. Note that most parameters were recovered with r>0.7, with the least faithfully recovered parameter being the state transition observation rate η^{tm} with r < 0.6. **B:** Correlations between fitted parameters. Note that only correlations between ∇_{pell2cue} and w^{mf} $(r = -0.54)$ in HB and between ∇_{pell2cue} and η^{tm} $(r = -0.57)$ are substantial.

bioRxiv preprint doi: https://doi.org/10.1101/2022.01.25.477716; this version posted January 26, 2022. 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-NC-ND 4.0 International license.

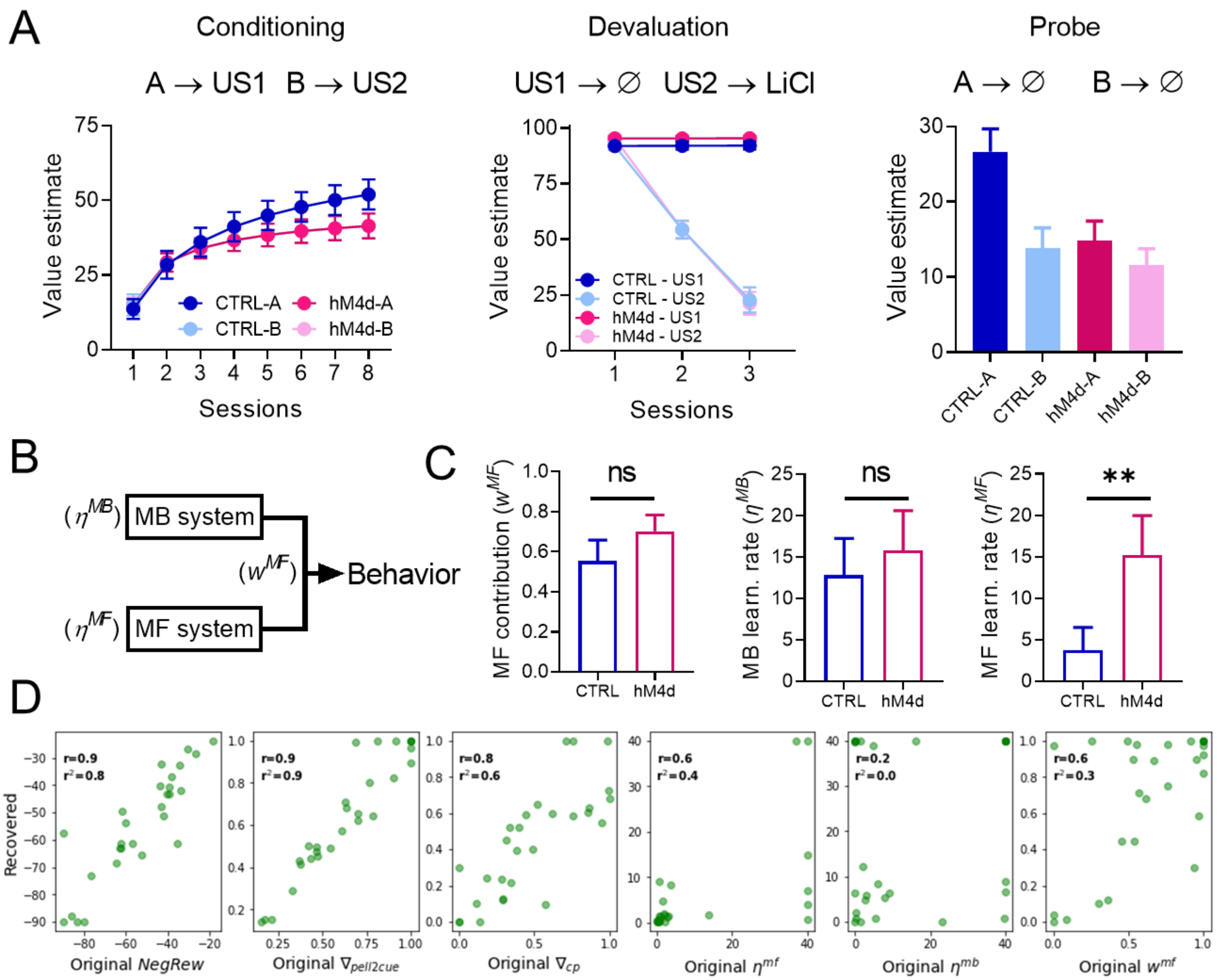

Figure S3. Data fitting with a reinforcement learning model that allows for a shift between model-based (MB) and model-free (MF) learning. (A) Model fit results for our MB vs MF reinforcement learning model. Note that it can also replicate our behavioral results well. (B) Schematic of the critical aspect of the model and the expected result: the observation rate for both the MB and MF systems, as well as the potential contribution of each to behavior, were free parameters, and we expected that the contribution of the MB system would be diminished, either by a reduced MB observation rate or an increase in the MF contribution. (C) Values of the critical observation rate-related parameters, namely the proportion of contribution of the MF (w^{mf}) system, the MF observation rate (η^{mf}), and the MB observation rate (η^{mb}) for both control and hM4d model fits. Note that instead of a reduction in MB learning or proportional contribution, only the MF observation rate was significantly higher in the hM4d group. See table S2 for detailed parameter comparisons. (D) Correlations between estimated and original parameters for the MB vs MF model. Note that parameter recovery of all critical observation rate-related parameters was not very faithful (r < 0.7). Data are represented as mean ± SEM. ***P*<0.01.