Emergence of rapid value inference through meta-reinforcement learning

Jaeeon Lee[1,2*], Jay A. Hennig[2,3,4,5], Vanessa Frelih[1,2], Samuel J. Gershman[2,3], Naoshige Uchida[1,2*]

[1]Department of Molecular and Cellular Biology, Harvard University, Cambridge, MA, USA

[2]Center for Brain Science, Harvard University, Cambridge, MA, USA

[3]Department of Psychology, Harvard University, Cambridge, MA, USA

[4]Department of Neuroscience, Baylor College of Medicine, Houston, TX, USA

[5]Neuroengineering Initiative, Rice University, Houston, TX, USA

[*]Correspondence: uchida@mcb.harvard.edu, jaylee@g.harard.edu

The ability to estimate the value associated with a specific stimulus or action is essential for adaptive behavior. Value can be updated either incrementally through experience or rapidly by inference based on latent environmental structure. Yet, how the brain implements and transitions between these modes of value computation remains unclear. To address this question, we examined the neuronal mechanisms underlying reversal learning. Mice were trained in an odor-outcome association task either with stable or dynamically changing contingencies. Mice trained on stable contingencies formed long-term value representations that depended on synaptic plasticity in the basolateral amygdala (BLA). In contrast, mice exposed to repeated reversals acquired the ability to infer values, independent from plasticity in BLA, enabling faster learning but with more rapid memory decay. Recurrent neural network models (RNNs) trained with continuous weight updates recapitulated this transition, shifting from plasticity-based to dynamics-based value computation. Neural activity in the BLA encoded both value and contextual information necessary for computing value based on latent task structure, similar to those found in the RNNs. Disrupting BLA activity before cue delivery preferentially impaired dynamics-based value updating. Furthermore, mice could learn distinct correlation structures that enabled structure-specific value inference. Together, these findings provide a mechanistic framework for fast value updates via inference, a core feature of intelligent behavior.

Main

Animals must continually estimate the value of sensory cues and actions to guide adaptive behavior. Reinforcement learning provides an algorithmic framework for this process[1–3], yet the neural mechanisms by which value is learned, stored, and flexibly updated remain unresolved. Classical theories emphasize incremental trial-and-error learning, in which changes in synaptic strength encode long-term value memories[4–11]. Indeed, plasticity within circuits such as the striatum and amygdala has been linked to the formation of stable appetitive and aversive associations.

In contrast, accumulating evidence suggests that animals can update value through inference by exploiting structural knowledge of the environment[12–21]. In RNNs, meta-reinforcement learning — the emergence of a fast reinforcement learning algorithm through a slower reinforcement learning algorithm — gives rise to inference-like behavior[22–24]. In these pre-trained models, value can be updated without online synaptic changes, through recurrent dynamics that encode hidden task states. Although both incremental learning and inference are observed behaviorally, the neural mechanisms supporting each — and how the brain transitions between them — remain elusive. A major challenge is that behavioral performance alone often cannot distinguish between plasticity-based and dynamics-based value updating, and the two mechanisms may coexist within the same circuit[25–27].

To address this, we developed two classical conditioning paradigms that differed in outcome stability. Combining behavior, electrophysiology, and computational modeling, we find that rapid value inference emerges through a gradual transition from synaptic plasticity-dependent learning to plasticity-independent value updating mediated by recurrent dynamics that encode task structure.

Timescale of value update/decay in stable vs dynamic task

We trained mice to perform a head-fixed classical conditioning task in which an odor cue (conditioned stimulus, CS+ or CS-) was followed by either a water reward or no reward (**Fig. 1a**). We used two versions of the task. In the stable task, the reward contingency was fixed, and mice had to initially learn a fixed value and maintain the value memory in the subsequent sessions (**Fig. 1b**, *top*). In the dynamic task, the reward contingency reversed every session with the reversal happening in the middle of each session (**Fig. 1b***, bottom*). Each session started with the reward contingency from the previous session so that mice had an incentive to remember previously learned values. We used anticipatory licking during the odor period as a proxy for value to quantify performance on both tasks. In the stable task, mice quickly learned to lick more to the CS+ odor than to the CS- odor from day 1 of training, reaching expert performance by day 3 (**Fig. 1c, Extended Data Fig. 1a, c**). In the dynamic task, mice gradually improved their performance over 12 days of training, after which expert mice reliably discriminated CS+ and CS- both before reversal (block 1) and after reversal (block 2), achieving comparable performance as in the expert in the stable task (**Fig. 1d,e, Extended Data Fig. 1b, c**).

To better understand how value updating at the beginning of the stable task and late stage of the dynamic task differed, we quantified the learning curve for CS+ or CS- lick rate (**Fig. 1f**). For the stable task, we plotted the CS+/CS- lick rate on the 1st day of the stable task or the 1st reversal session (1st dynamic task after being trained on the stable task). This was compared to the CS+/CS- lick rate aligned to the reversal point in expert mice in the dynamic task. For both positive value update (+Δvalue) and negative value update (–Δvalue), learning rate was much faster in the dynamic vs stable task (for positive update, τ_{stable}=80.5 trials; τ_{dynamic}=2.4 trials; for negative update, τ_{stable}=52.3 trials; τ_{dynamic}=8.1 trials). Overall, these results suggest that as mice transition from a stable to a dynamic environment, the timescale of value updates, as measured by the conditioned response (e.g. anticipatory licking), gets faster by an order of magnitude.

One interesting possibility is that mice transition from a plasticity- to a dynamics-based value updating strategy as they move from the stable to dynamic environment. Dynamics-based value updating may require mice to maintain information about value via persistent activity, which might be prone to temporal degradation akin to working memory[28–30]. We thus reasoned that value memory might degrade at a distinct timescale in the stable vs. dynamic tasks. To test this hypothesis, we first quantified how fast value memory degrades over time in the stable or dynamic task. In both tasks, the reward contingency at the beginning of each session was the same as that of the end of the previous session (**Fig. 1g**). Thus, we asked if mice correctly discriminated CS+ and CS- odors on the very first trial in each session. Mice were fully trained on either the stable task (5 days of training) or dynamic task (12 days of training) and then tested for value memory by quantifying the number of licks during the first trial of each cue. In the stable task, mice performed correctly from the first trial by licking to the first CS+ and not the CS- (**Fig. 1g**, stable task). However, in the dynamic task, mice started the session by licking to both CS+ and CS- even though they had discriminated CS+ and CS- at the end of the previous session (**Fig. 1g**, dynamic task). Mice then quickly learned to suppress licking to the CS-, allowing them to still maintain high discriminability between CS+ and CS- throughout the block (**Extended Data Fig. 1d**). To quantify this "forgetting" effect more systematically, we computed a discrimination index for each session, with index=1 being perfect discrimination between first CS+ and first CS-, index=0 being no discrimination, and index=-1 being the flipped discrimination (see Methods). We also varied the duration of the break between each session of the stable task by pausing behavioral training for up to 8 days. In the stable task, the discrimination index was close to 1 even with 8 days of break whereas in the dynamic task, the index was close to zero even after a 1-day break (**Fig. 1h**). Overall, these results suggest that value memory in the stable task persists over 8 days, whereas value memory degrades to chance level after only 1 day in the dynamic task.

To further quantify the timescale at which dynamic value memory degrades, we introduced an inter-trial interval (ITI) that was longer than the average duration that mice were initially trained on (ITI=37.5-300 sec) in the middle of each session in both the stable and dynamic task (two long-ITI trials per session) (**Fig. 1i**, *left*). These extended ITIs caused a similar phenotype as the break between sessions: value memory decayed to chance level in the dynamic task after 300 seconds but not in the stable task (**Fig. 1i**, *right,* **Extended Data Fig. 1e**). The discrimination index was significantly lower for dynamic vs stable tasks for duration of 150 and 300 seconds. As an alternative method to measure value, we expressed a dopamine sensor (GRAB^{DA3m}) in the ventral striatum and quantified the dopamine cue response as a measure for value (**Extended Data Fig. 1f-h**). The discrimination index for the dopamine response was similar to the discrimination index for the CS lick rate, suggesting that both behavioral and neural readout of value display a similar forgetting timescale between the two tasks.

RNN model with continual plasticity

The above results suggest that mice might initially store value information in synaptic weights when the environment is stable, but repeated exposure to value reversals might cause a transition from a plasticity-based value update to a dynamics-based value update, allowing faster value updating at the expense of being more forgetful. To better understand how this transition might occur mechanistically, we took a computational modelling approach. Previous works using the temporal difference (TD) learning algorithm have been successful at explaining various features of dopamine responses and value learning[4,31–33]. However, most previous applications of TD to animal learning *a priori* assumed a specific state representation that is fixed and thus does not model the emergence of the state representation itself. On the other hand, more recent work has shown that TD learning models equipped with RNNs can learn useful representations directly (e.g. beliefs about hidden states), mirroring activity seen in actual neural recordings[22,25,34,35]. However, many of these models are trained offline. This implies that plasticity in these models cannot play a causal role in updating value online—in stark contrast to classical TD modeling approaches, where value is updated online but exclusively via plasticity. Thus, to create a biologically realistic model of learning that is both agnostic about state representations and also displays online plasticity updates, we trained RNNs with TD learning in an online fashion (**Fig. 2a**, see Methods), where at every timestep, the RNN's weights were updated based on the recent history of inputs. Unlike previous methods using RNNs to model value learning, weights in the RNN were never frozen, allowing for a continual interaction between recurrent dynamics and plasticity.

We trained RNNs on either the stable task or dynamic task (**Fig. 2b-c, Extended Data Fig. 2a-c-d**; see Methods). The RNNs' value readouts could reliably discriminate between cue A and cue B (**Fig. 2b**). Interestingly, in the dynamic task, the learning rate for updating value displayed an abrupt transition from a slow update regime to a fast update regime (**Fig. 2b**, right). We quantified the learning rate for positive or negative value updates in the stable and dynamic tasks (**Fig. 2c**). Similar to the mouse behavioral data (**Fig. 1f**), RNNs displayed faster value updates in the dynamic task compared to the first reversal in the stable task (**Fig. 2c**). To better understand the mechanisms driving value updates in these two conditions, we manipulated plasticity in the RNNs by setting the learning rate of the RNN weight update to zero (**Fig. 2d**, see Methods). Without plasticity, RNNs were completely impaired at updating value in the stable task, whereas RNNs were still able to update value in the dynamic task (**Fig. 2d**, *right*). We quantified the difference between the value readouts of CS+ and CS- cues in RNNs with plasticity ($\Delta w \neq 0$) or without plasticity ($\Delta w=0$), for RNNs trained on either the stable or dynamic task (**Fig. 2e**). Without plasticity, the difference between predicted values for CS+ and CS- cues decreased in the stable task but not in the dynamic task. Overall, these findings suggest that RNNs with online weight updating initially use plasticity to learn value in the stable task, but transition to a plasticity-independent mechanism in the dynamic task.

What could be the mechanism driving value updating in the dynamic task? To answer this question, we applied principal component analysis (PCA) to the activity in the RNNs (see Methods) and plotted the neural-state space trajectories with an additional axis representing the value predicted by the RNNs (**Fig. 2f, g**). In the stable task, neural trajectories for CS+ gradually changed so that the RNN's response to the cue (green point) moved upwards towards higher value, representing plasticity-driven value updating (cue A, **Fig. 2f**). In the dynamic task, neural trajectories for an expert RNN were segregated by cue type and block type, with PC1 encoding cue type and PC2 encoding block type (**Fig. 2g**, *left*). Information about block identity, or context information, could potentially be used in the RNN to decode value information for each block (**Fig. 2g**, *right*). To see if the contextual information present in the RNN activity was driving value computation in the RNNs, we computed the context axis, which was defined as the linear discriminant axis that best separated context information in an expert RNN (see Methods). We then projected the hidden units' activity during the ITI onto the context axis ($\text{context}^{\text{proj}}$), and plotted it along the value readout for cue A and cue B (**Fig. 2h**). When value was updated slowly during the initial phase of learning, $\text{context}^{\text{proj}}$ did not discriminate block identity very well. However, as the RNN transitioned into a faster value update regime, $\text{context}^{\text{proj}}$ started to discriminate block identity. We calculated the Spearman correlation between the difference in value (value diff.) on the last trial to cue A and cue B, and $\text{context}^{\text{proj}}$ in either the naïve or expert RNN (**Fig. 2i**). The correlation was larger in expert RNNs compared to naïve RNNs (**Fig. 2j**, *left*). Furthermore, the activity in different context (context A or B) projected onto the discriminant axis, became more distinct in expert mice, suggesting that context information became more discrete and separable, and potentially indicating the emergence of fixed points corresponding to each context. Overall, these results suggest that in the dynamic task, RNNs gradually develop a representation of block identity (e.g., using fixed points) which allow the RNNs to rapidly update the value of each cue across blocks through neural dynamics (i.e. without plasticity).

Lastly, we simulated the effect of introducing a long ITI during the stable or dynamic task. Given that context information encoded in the hidden units' activity is potentially important for computing value in the dynamic task, we reasoned that a long ITI might cause a drift in the activity in the RNN, resulting in an ITI duration-dependent change in value discrimination. Consistent with this prediction, a long ITI caused a change in $\text{context}^{\text{proj}}$ and a subsequent change in value readout in the RNN (**Fig. 2k**). We quantified the discrimination index similarly to experimental data in

Fig. 1h, j. We found that a long ITI caused a duration-dependent change in discrimination index only in the dynamic task, recapitulating experimental data (**Fig. 1j**). This effect could be explained by the magnitude of the drift along the context axis, with larger drift causing a larger change in value difference (**Fig. 2m**). Overall, these results suggest that contextual information, represented by the population activity, is prone to temporal degradation, thus causing a time-dependent degradation of value memory in the dynamic task.

The role of plasticity and activity in the BLA

Computational modelling with RNNs suggested that a transition from plasticity to dynamics-based value update could explain the experimental data. This implies that blocking synaptic plasticity in the region of the brain responsible for value updating should impair performance in the stable task on day 1, but not performance in the dynamic task after being fully trained. We tested this prediction by targeting the BLA, a brain region that has been previously shown to undergo synaptic plasticity in both appetitive and aversive tasks as well as showing activity correlated with conditioned responding[7,11,36–39]. To block synaptic plasticity acutely, we injected a CaMKII blocker (KN-93) locally in the BLA and tested the performance on either day 1 of the stable task or on expert stage of the dynamic task (**Fig. 3a, Extended Data Fig. 1a-e**). KN-93 has been shown to effectively block synaptic plasticity in slice, and local infusion *in vivo* has been shown to cause behavioral effects consistent with an impairment of synaptic plasticity[40–44]. KN-93 infusion in the BLA significantly decreased the difference in anticipatory licking between CS+ and CS- trials in the stable task compared to saline infusion (**Fig. 3b**, *left*). In stark contrast, the same manipulation had no effect in the dynamic task (**Fig. 3b**, *right*). These results suggest that BLA plasticity is necessary to initially update value in the stable task, but becomes dispensable in the dynamic task.

One alternative explanation for the dissociable role of BLA plasticity in the stable vs dynamic tasks is that BLA might become disengaged in the dynamic task, with another brain region taking over the role of BLA. Thus, plasticity in another brain region other than BLA might still be responsible for updating value in the dynamic task (**Extended Data Fig. 3f**). This alternative model would predict that BLA activity is only necessary to perform the stable task and not the dynamic task. To test if this was true, we acutely inactivated BLA activity. We generated an emx1-Cre × gtACR1 mice, in which the inhibitory opsin is expressed in a Cre-dependent manner in the excitatory neurons throughout the brain including BLA. BLA specificity was achieved by implanting an optical fiber just above BLA in emx1-cre × gtACR1 mice. Mice were trained in either the stable or dynamic task, after which BLA was inactivated on 15% of all trials during the cue period for 3 seconds (**Fig. 3c**). Inactivating BLA neurons impaired performance in both stable and dynamic tasks, by increasing the CS- licks, resulting in poorer discrimination of CS+/CS- (**Fig. 3d**). Thus, BLA activity is still necessary for performing the dynamic task, despite plasticity in BLA becoming dispensable.

Value coding in the amygdala

To better understand the nature of value coding in the BLA, we performed acute high-density electrophysiological recording using Neuropixels probes in the BLA (**Fig. 4a, b**). We modified the original task so that both stable and dynamic value could be measured in the same recording session (**Fig. 4a**). The hybrid task consisted of 3 odor cues, two of which were stable odor cues (odor A=reward, odor B=no reward), and one of which was a dynamic odor cue (odor C=reward or no reward depending on block). The value of odor C changed throughout the session across 4 blocks. In all sessions, reward contingency for odor C started with the same condition as the previous session to ensure continuity of value (**Fig. 4a**, *bottom*). Consistent with the timescale of value memory decay described previously (**Fig. 1h**), mice forgot the value of odor C between sessions, as indicated by mice always licking to odor C at the beginning of the session (**Extended Data Fig. 4**). This behavior was not present in odor B, suggesting that value memory for stable and dynamic odors had distinct timescales of memory decay in the hybrid task.

Analysis of neural recording data revealed that BLA units encoded both stable and dynamic values, with many units encoding both (**Fig. 4c**). An example stable value coding (SV) unit in the BLA consistently differentiated odor A and odor B across blocks (**Fig. 4c**, top). An example dynamic value coding (DV) unit in the BLA responded to odor C differentially depending on its value in that block (**Fig. 4c**, middle). An example unit in the BLA encoded both SV and DV (SVDV, **Fig. 4c**, bottom). Given that Neuropixels recording allowed wide sampling of brain regions surrounding BLA, we analyzed the fraction of SV, DV and SVDV units across all recorded brain regions (**Fig. 4d**). SV, DV and SVDV units were enriched in the amygdala, especially in the BLA and BMA (BLA: SV fraction=35%, DV fraction=13%, SVDV fraction=8%; BMA: SV fraction=34%, DV fraction=9%, SVDV fraction=6%). We defined the polarity of stable value or dynamic value by whether firing rate was higher for the rewarded cue or not (see Methods). We plotted the mean firing rate across all positive SV, negative SV, positive DV or negative DV in the amygdala (**Fig. 4e**). Interestingly, positive SV and DV units tended to fire phasically to both odors (**Fig. 4e**, left panels). In contrast, negative SV and DV units tended to be bidirectionally modulated relative to baseline (**Fig. 4e**, right panels). To see if the polarity of SV and DV was congruent in SVDV units, we plotted the SV selectivity vs DV selectivity for all SVDV units (**Fig. 4f**, left). Selectivity was defined as the difference in firing rate between the rewarded cue/block and non-rewarded cue/block (SV selectivity=odor A- odor B; DV selectivity=odorC$^{\text{reward}}$ – odorC$^{\text{noreward}}$). We found a strong positive correlation between SV selectivity and DV selectivity (**Fig. 4f**). The fraction of units that had congruent polarity for SV and DV was higher than units that had incongruent polarity (**Fig. 4f**, right). Overall, these results suggest that amygdala contains neurons that encode both SV and DV in a congruent manner.

Context coding in the amygdala

BLA can compute and store stable value using plasticity (**Fig. 3a, b**) but requires context information to compute dynamic value (**Fig. 2h-j**). We considered two hypotheses for how BLA could acquire dynamic value. In one scenario, dynamic value is computed outside BLA and inherited by BLA (model 1, **Extended Data Fig. 5a**). In another scenario, BLA locally computes dynamic value using context information within BLA (model 2, **Extended Data Fig. 5a**). To distinguish between these two scenarios, we looked for cells in the BLA that could differentiate block types during the ITI. To our surprise, we found many units in the BLA that contained context information during the ITI period (**Fig. 5, Extended Data Fig5. b)**. An example unit in the BLA differentiated context by firing higher during the ITI period of non-rewarded blocks (**Fig. 5a-b**). This unit's firing rate was negatively correlated with the odor C CS licks, suggesting that ITI firing rate was predictive of upcoming anticipatory licking or value for odor C (negative context unit). Out of all context units in the amygdala, we found that a large fraction of them also encoded stable and dynamic value, suggesting that context and value information might be multiplexed in the amygdala (**Extended Data Fig. 5c**). We analyzed all the brain region recorded and found that BLA and BMA were enriched with context coding units, with negative context units being more dominant than positive context units (BLA: 5.7%; BMA: 4.2%) (**Fig. 5c**). To test if the ITI firing rate was predictive of the anticipatory CS licks to odor C, we computed the Spearman's correlation coefficient between odor C CS licks and the ITI firing rate before cue onset for each neuron (**Fig.**

5d). Importantly, we restricted our analysis to contain only one block type to avoid circular analysis (see Methods). We found that positive or negative context coding units had a statistically significant correlation coefficient, suggesting that the ITI activity of context units predicts upcoming behavioral choice for dynamic odor C (**Fig. 5e**). Lastly, given that context coding is confounded by reward rate coding in this task, we asked if ITI firing rate was updated in a manner consistent with pure context coding or pure reward rate coding. We reasoned that if context coding is prevalent, then ITI firing rate should only be updated after dynamic odor C, and not after stable odors A or B. However, if reward rate coding is prevalent, then ITI firing rate might be updated similarly to all outcomes regardless of cue type (**Extended Data Fig. 5d**). Context update analysis revealed that change in ITI firing rate was more consistent with context coding than reward rate coding (**Fig. 5f**). This was especially true for negative context units, which was most of the context coding units. To see if this context coding was causal to value computed using dynamics, we trained emx1-Cre × gtACR1 mice on either stable or dynamic mice, after which BLA was bilaterally inactivated during the ITI (**Fig. 5g**). We reasoned that if context information during the ITI is important for computing dynamic value, disrupting activity in this period should impair performance during the dynamic task, but not during the stable task where value is not computed using dynamics. BLA inactivation during the ITI did not change the performance in the stable task but increased CS- licks in the dynamic task (**Fig. 5h**). When we computed the difference between the mean CS+ licks and CS- licks, this difference was significantly reduced in the dynamic task but not in the stable task (**Fig. 5i**). Overall, these results suggest that BLA contains context information which not only predicts upcoming behavior, but is also necessary for performing in the dynamic task.

Dynamics allow value inference

We have shown so far that value updating via dynamics can be fast but forgetful compared to value updating via plasticity. Another advantage of using dynamics over plasticity is that dynamics-based value updates support value inference, which is the ability to update value without direct experience. This is because RNNs naturally learn the structure of the task and encode the hidden states of the task (**Fig. 2**). Given that in the dynamic task, there are essentially two hidden states corresponding to each block, RNNs using dynamics can infer value for one cue without direct experience (counterfactual learning; **Fig. 6a**). To look for evidence that our RNNs can perform value inference, we designed a paradigm in which, after reversal, we only presented one cue type for up to 20 trials, after which we presented the opposite cue (probe cue) for the first time (F**ig. 6b**) An agent that understands the structure of the task (i.e. anti-correlation in the values predicted by odor A and B) would be able to infer the change in value for the probe cue, whereas an agent that does not understand the structure of the task might not infer any change and start from the pre-reversal point. We tested this in both RNNs and mice (**Fig. 6c-g, EDF 6a)**. In RNNs, value inference only emerged in expert RNNs trained on the dynamic task (**Fig. 6c**). This is consistent with the idea that a naïve RNN uses plasticity, whereas an expert RNN that uses dynamics alone can infer value using its learned dynamics. Similarly in mice, we found that naïve mice undergoing reversal for the first time failed to infer a change in value, whereas expert mice fully trained on the dynamic task could infer a change in value (**Fig. 6d**). This effect was consistent across many RNNs and mice as measured by CS licks (**Fig. 6e, f**, **EDF 6c, d**). If cue-evoked dopamine reports value, then we should also expect dopamine signals to reflect inferred value change. Consistent with previous studies[14,16,17], we also found that dopamine signals reflected inference (**Fig. 6f**, *right,* **EDF 6b**). We also found that the magnitude of the inferred change in value (ΔCS licks) parametrically varied as a function of the number of opposite trials, with a larger number of trials leading to bigger change in ΔCS licks (**Fig. 6g, EDF 6c, d**).

Inference has been mostly studied using tasks implementing anti-correlated values between two options[12–14,19]. To test if mice can learn distinct structures to guide structure-specific inference, we implemented three correlation structures of odor values. We trained separate cohorts of mice in environments in which a pair of odor-value associations are either anti-correlated, correlated, or independent. We then tested for the presence of inference in a similar fashion as above (**Fig. 6g**). Indeed, mice displayed signatures of inference consistent with the pre-exposed correlation structure as measured based on anticipatory licks and cue-evoked dopamine responses (**Fig. 6h, i**). For instance, when the value of odor A changes from positive to zero, mice trained in the anti-correlated structure increased the value of odor B without directly experiencing the new contingency for odor B, whereas mice trained in the correlated structure decreased the value of odor B. Mice trained in the independent structure exhibited a negligible sign of inference. Overall, these results suggest that a transition from plasticity- to dynamics-based value update allows rapid value inference to emerge and is specific to the learned correlated structure.

Discussion

Through a combination of behavioral, circuit-level, and computational approaches, we show that value computation can transition from a synaptic plasticity-based to a recurrent dynamics-based mechanism allowing value inference to emerge in a dynamic environment. Because contextual representations were maintained during the ITI, the outcome of one cue updated the inferred value of both cues simultaneously. Mice could also learn distinct correlation structures, suggesting inference can be adaptive. Overall, our work provides a mechanistic framework for understanding how rapid inference emerges in the brain.

Computationally, our results extend prior models showing that recurrent network dynamics can encode value without synaptic change[23,25,26,45]. Previous studies typically used an offline-learning rule in which the weights are only updated outside the task and then fixed after convergence. Thus, these networks assume that plasticity does not play a role for incremental online improvement in performance. Our model incorporates continuous online plasticity during task performance, allowing a much more biologically realistic interaction between plasticity and dynamics. Moreover, we show that a single learning rule—truncated backpropagation through time (TBPTT)—can produce both plasticity-based and dynamics-based value updates, depending on the timescale of the learning window. Shorter windows favor value update based on plasticity, while longer windows enable meta-reinforcement learning through emergent recurrent dynamics (**Extended Data Fig. 2f, g**). One interesting possibility is that distinct brain areas may implement the same underlying rule with different effective timescales, enabling diverse computational functions to emerge in different brain regions[46–48].

Behaviorally, we demonstrate that mice exploit environmental regularities in the task to infer value, consistent with prior work showing across species that animals display inference behavior and that regions like the hippocampus and OFC contribute to inference-based decision-making[12–14,17,31,32,49,50]. Our findings highlight the amygdala's contribution to context-dependent inference[51], raising the question of whether contextual representations are inherited from inputs such as the ventral hippocampus and OFC, or computed locally within the amygdala. Disentangling these possibilities will be essential for understanding how inference unfolds across distributed neural circuits.

Finally, our results reveal a fundamental tradeoff between stability and flexibility in value computation. Dynamics-based value representations enable rapid value update using inference at the cost of degrading over time, whereas plasticity-based representations provide stable long-term storage at the cost of being slow and inflexible. Given the further cost of having to maintain information in the persistent activity, dynamics-based value update might only emerge when the benefits of faster value update (e.g. environment is dynamic) is high. Elucidating the exact conditions under which these two modes of value update dominate will be crucial in the future.

References

1. Sutton, R. S. & Barto, A. G. *Reinforcement Learning: An Introduction*. (A Bradford Book, Cambridge, Mass, 1998).
2. Botvinick, M., Wang, J. X., Dabney, W., Miller, K. J. & Kurth-Nelson, Z. Deep Reinforcement Learning and Its Neuroscientific Implications. *Neuron* **107**, 603–616 (2020).
3. Lee, D., Seo, H. & Jung, M. W. Neural Basis of Reinforcement Learning and Decision Making. *Annual Review of Neuroscience* **35**, 287–308 (2012).
4. Schultz, W., Dayan, P. & Montague, P. R. A Neural Substrate of Prediction and Reward. *Science* **275**, 1593–1599 (1997).
5. Reynolds, J. N. J., Hyland, B. I. & Wickens, J. R. A cellular mechanism of reward-related learning. *Nature* **413**, 67–70 (2001).
6. Hong, S. & Hikosaka, O. Dopamine-Mediated Learning and Switching in Cortico-Striatal Circuit Explain Behavioral Changes in Reinforcement Learning. *Front. Behav. Neurosci.* **5**, (2011).
7. Tye, K. M., Stuber, G. D., de Ridder, B., Bonci, A. & Janak, P. H. Rapid strengthening of thalamo-amygdala synapses mediates cue–reward learning. *Nature* **453**, 1253–1257 (2008).
8. Yamaguchi, K. *et al.* The minimal behavioral time window for reward conditioning in the nucleus accumbens of mice. 43.
9. Xiong, Q., Znamenskiy, P. & Zador, A. M. Selective corticostriatal plasticity during acquisition of an auditory discrimination task. *Nature* **521**, 348–351 (2015).
10. Tsutsui-Kimura, I. *et al.* Dopamine in the tail of the striatum facilitates avoidance in threat–reward conflicts. *Nat Neurosci* **28**, 795–810 (2025).
11. Rogan, M. T., Stäubli, U. V. & LeDoux, J. E. Fear conditioning induces associative long-term potentiation in the amygdala. *Nature* **390**, 604–607 (1997).
12. Mishchanchuk, K. *et al.* Hidden state inference requires abstract contextual representations in ventral hippocampus. Preprint at https://doi.org/10.1101/2024.05.17.594673 (2024).
13. Vertechi, P. *et al.* Inference-Based Decisions in a Hidden State Foraging Task: Differential Contributions of Prefrontal Cortical Areas. *Neuron* **106**, 166-176.e6 (2020).
14. Bromberg-Martin, E. S., Matsumoto, M., Hong, S. & Hikosaka, O. A Pallidus-Habenula-Dopamine Pathway Signals Inferred Stimulus Values. *Journal of Neurophysiology* **104**, 1068–1076 (2010).

15. Qü, A. J. *et al.* Nucleus accumbens dopamine release reflects Bayesian inference during instrumental learning. *PLOS Computational Biology* **21**, e1013226 (2025).

16. Blanco-Pozo, M., Akam, T. & Walton, M. E. Dopamine-independent effect of rewards on choices through hidden-state inference. *Nat Neurosci* **27**, 286–297 (2024).

17. Takahashi, Y. K., Stalnaker, T. A., Roesch, M. R. & Schoenbaum, G. Effects of inference on dopaminergic prediction errors depend on orbitofrontal processing. *Behavioral Neuroscience* **131**, 127–134 (2017).

18. Barron, H. C. *et al.* Neuronal Computation Underlying Inferential Reasoning in Humans and Mice. *Cell* **183**, 228-243.e21 (2020).

19. Courellis, H. S. *et al.* Abstract representations emerge in human hippocampal neurons during inference. *Nature* **632**, 841–849 (2024).

20. Jang, A. I. *et al.* The Role of Frontal Cortical and Medial-Temporal Lobe Brain Areas in Learning a Bayesian Prior Belief on Reversals. *J. Neurosci.* **35**, 11751–11760 (2015).

21. Costa, V. D., Tran, V. L., Turchi, J. & Averbeck, B. B. Reversal Learning and Dopamine: A Bayesian Perspective. *J. Neurosci.* **35**, 2407–2416 (2015).

22. Hattori, R. *et al.* Meta-reinforcement learning via orbitofrontal cortex. *Nat Neurosci* **26**, 2182–2191 (2023).

23. Wang, J. X. *et al.* Prefrontal cortex as a meta-reinforcement learning system. *Nat Neurosci* **21**, 860–868 (2018).

24. Kim, C. M., Chow, C. C. & Averbeck, B. B. Neural dynamics of reversal learning in the prefrontal cortex and recurrent neural networks. *eLife* **13**, RP103660 (2025).

25. Parker, N. F. *et al.* Choice-selective sequences dominate in cortical relative to thalamic inputs to NAc to support reinforcement learning. *Cell Reports* **39**, (2022).

26. Pereira-Obilinovic, U., Hou, H., Svoboda, K. & Wang, X.-J. Brain mechanism of foraging: Reward-dependent synaptic plasticity versus neural integration of values. *Proceedings of the National Academy of Sciences* **121**, e2318521121 (2024).

27. Durstewitz, D., Averbeck, B. & Koppe, G. What neuroscience can tell AI about learning in continuously changing environments. *Nat Mach Intell* **7**, 1897–1912 (2025).

28. Bae, J. W. *et al.* Parallel processing of working memory and temporal information by distinct types of cortical projection neurons. *Nat Commun* **12**, 4352 (2021).

29. Zhang, X. *et al.* Active information maintenance in working memory by a sensory cortex. *eLife* **8**, e43191 (2019).

30. Bolkan, S. S. *et al.* Thalamic projections sustain prefrontal activity during working memory maintenance. *Nat Neurosci* **20**, 987–996 (2017).

31. Babayan, B. M., Uchida, N. & Gershman, S. J. Belief state representation in the dopamine system. *Nat Commun* **9**, 1891 (2018).

32. Starkweather, C. K., Babayan, B. M., Uchida, N. & Gershman, S. J. Dopamine reward prediction errors reflect hidden-state inference across time. *Nat Neurosci* **20**, 581–589 (2017).

33. Gershman, S. J. *et al.* Explaining dopamine through prediction errors and beyond. *Nat Neurosci* **27**, 1645–1655 (2024).

34. Hennig, J. A. *et al.* Emergence of belief-like representations through reinforcement learning. *PLOS Computational Biology* **19**, e1011067 (2023).

35. Qian, L. *et al.* Prospective contingency explains behavior and dopamine signals during associative learning. *Nat Neurosci* **28**, 1280–1292 (2025).

36. Beyeler, A. *et al.* Organization of Valence-Encoding and Projection-Defined Neurons in the Basolateral Amygdala. *Cell Reports* **22**, 905–918 (2018).

37. Paton, J. J., Belova, M. A., Morrison, S. E. & Salzman, C. D. The primate amygdala represents the positive and negative value of visual stimuli during learning. *Nature* **439**, 865–870 (2006).

38. Zhang, X. & Li, B. Population coding of valence in the basolateral amygdala. *Nat Commun* **9**, 5195 (2018).

39. Zhang, X. *et al.* Genetically identified amygdala–striatal circuits for valence-specific behaviors. *Nat Neurosci* **24**, 1586–1600 (2021).

40. Yagishita, S. *et al.* A critical time window for dopamine actions on the structural plasticity of dendritic spines. *Science* https://doi.org/10.1126/science.1255514 (2014) doi:10.1126/science.1255514.

41. Radiske, A., de Castro, C. M., Rossato, J. I., Gonzalez, M. C. & Cammarota, M. Hippocampal CaMKII inhibition induces reactivation-dependent amnesia for extinction memory and causes fear relapse. *Sci Rep* **13**, 21712 (2023).

42. Adler, A., Zhao, R., Shin, M. E., Yasuda, R. & Gan, W.-B. Somatostatin-Expressing Interneurons Enable and Maintain Learning-Dependent Sequential Activation of Pyramidal Neurons. *Neuron* **102**, 202-216.e7 (2019).
43. Cichon, J. & Gan, W.-B. Branch-specific dendritic Ca2+ spikes cause persistent synaptic plasticity. *Nature* **520**, 180–185 (2015).
44. Radiske, A. *et al.* Avoidance memory requires CaMKII activity to persist after recall. *Molecular Brain* **14**, 167 (2021).
45. Kim, C. M., Chow, C. C. & Averbeck, B. B. Neural dynamics of reversal learning in the prefrontal cortex and recurrent neural networks. *eLife* **13**, (2025).
46. Mohebi, A., Wei, W., Pelattini, L., Kim, K. & Berke, J. D. Dopamine transients follow a striatal gradient of reward time horizons. *Nat Neurosci* **27**, 737–746 (2024).
47. Murray, J. D. *et al.* A hierarchy of intrinsic timescales across primate cortex. *Nat Neurosci* **17**, 1661–1663 (2014).
48. Song, M. *et al.* Hierarchical gradients of multiple timescales in the mammalian forebrain. *Proceedings of the National Academy of Sciences* **121**, e2415695121 (2024).
49. Hampton, A. N., Bossaerts, P. & O'Doherty, J. P. The Role of the Ventromedial Prefrontal Cortex in Abstract State-Based Inference during Decision Making in Humans. *J. Neurosci.* **26**, 8360–8367 (2006).
50. Baram, A. B., Muller, T. H., Nili, H., Garvert, M. M. & Behrens, T. E. J. Entorhinal and ventromedial prefrontal cortices abstract and generalize the structure of reinforcement learning problems. *Neuron* **109**, 713-723.e7 (2021).
51. Prévost, C., McNamee, D., Jessup, R. K., Bossaerts, P. & O'Doherty, J. P. Evidence for Model-based Computations in the Human Amygdala during Pavlovian Conditioning. *PLOS Computational Biology* **9**, e1002918 (2013).
52. Lee, J. & Sabatini, B. L. Striatal indirect pathway mediates action switching via modulation of collicular dynamics. *bioRxiv* 2020.10.01.319574 (2020) doi:10.1101/2020.10.01.319574.
53. Sumi, M. *et al.* The newly synthesized selective Ca2+calmodulin dependent protein kinase II inhibitor KN-93 reduces dopamine contents in PC12h cells. *Biochemical and Biophysical Research Communications* **181**, 968–975 (1991).
54. Werbos, P. J. Backpropagation through time: what it does and how to do it. *Proceedings of the IEEE* **78**, 1550–1560 (1990).

55. Williams, R. J. & Zipser, D. Gradient-based learning algorithms for recurrent networks and their computational complexity. in *Backpropagation: theory, architectures, and applications* 433–486 (L. Erlbaum Associates Inc., USA, 1995).

Methods

Experimental procedures

Animals. A total of 48 wild type (WT) C57BL/6J mice (Jackson Laboratory, male and female) were used in the experiments. For optogenetic inhibition of BLA, we crossed a cre-dependent gtACR1 reporter mouse (JAX:033089) with emx1-Cre mice (JAX: 005628) (n=5 mice) to label the excitatory populations within BLA. We used mice heterozygous for both alleles.

Animals were housed on a 12-h dark–12-h light cycle and performed the task at the same time each day (±1 h), during the dark period. Ambient temperature was kept at 75 ± 5 °F, and humidity was kept below 50%. Animals were group-housed (2–5 animals per cage) until surgery, then individually housed throughout the experiment. All procedures were performed in accordance with the US National Institutes of Health Guide for the Care and Use of Laboratory Animals and approved by the Harvard Institutional Animal Care and Use Committee.

Surgeries. All surgical procedures were conducted under aseptic conditions. Mice (older than 8 weeks) were anesthetized with isoflurane (3.5% for induction, followed by 1–2% for maintenance at 1 L min^{-1}). A local anesthetic (2% lidocaine) was administered subcutaneously at the incision site. Analgesia was provided with buprenorphine (0.1 mg kg^{-1}, i.p.) pre-operatively and ketoprofen (5 mg kg^{-1}, i.p.) for two days post-operatively. After leveling, cleaning, and drying the skull, a custom-made titanium head plate was attached using adhesive cement (C&B Metabond, Parkell). For all injections, the solution (virus or KN-93) was loaded into a pulled glass pipette (5-000-1001-X, Drummond) backfilled with mineral oil and fitted with a plunger. A small craniotomy (<1 mm diameter) was made using a dental drill, and the pipette assembly was mounted on a stereotaxic holder, lowered to the target coordinates, and injected slowly (~100 nL min^{-1}) to minimize tissue damage (MO-10, Narishige). After each injection, the pipette was left in place for at least 5 min to allow diffusion before being raised to the next site or withdrawn from the brain. Target coordinates for the brain regions were: ventral striatum (VS): 1.0/1.5/3.8mm, BLA: -1.65/3.3/4.2mm, (anterior-posterior/medial-lateral/dorsal-ventral; coordinates are relative to bregma, and DV relative to surface of the brain). Post surgery, mice were allowed to recover for at least 2 weeks for beginning the experiment.

Viruses. To record dopamine using fiber photometry, we expressed AAVs encoding the green dopamine sensor GRAB^{DA3m} (AAV-hsyn-DA3m(h-D05), WZ Biosciences, 300nl concentration=5×10^{12} gc/mL) in the left hemisphere VS of WT mice.

Behavioral setup. All behavioral experiments took place inside custom-built enclosed behavioral box in which head-fixed mice could stand on a fixed running wheel. The frame of the behavioral box was built using aluminum frames (McMaster) and the walls were made using black hardboard (Thorlabs, TB4). Mice had access to a water spout which delivered artificially sweetened water (Acesulfame Potassium powder dissolved in water at 6g/L; Prescribed For Life) for water reward, and an odor spout which delivered the odor cures for classical conditioning. Behavioral events were controlled (and licking was monitored) using custom-written software in MATLAB (Mathworks) and the Bpod library (Sanworks) interfacing with the Bpod state machine (1024 and 1027, Sanworks), valve module (1015, Sanworks) and port interface board (1020, Sanworks)/water valve (LHDA1233115H, Lee Company) assembly. Odors were delivered using a custom olfactometer, which directed air through one of eight solenoid valves (LHDA1221111H, Lee Company) mounted on a manifold (LFMX0510528B, Lee Company). Each odor was dissolved in mineral oil at 10% dilution, and 30 µL of diluted odor solution was applied to a syringe filter (2.7 µm pore, 13 mm diameter; 6823-1327, Whatman). Wall air was passed through a hydrocarbon filter (HT200-4, Agilent Technologies) and split into a 100 mL min^{-1} odor stream and 900 mL min^{-1} carrier stream using analogue flowmeters (MFLX32460- 40 and MFLX32460-42, Cole-Parmer), which were recombined at the odor manifold before being delivered to the nose of the mouse. During photometry experiments, licking was monitored using closed-loop circuit similar to previously described method[52]. During *in vivo* electrophysiology, an infrared emitter–photodiode. Infrared method is less prone to electrical artifact, and thus more suitable for *in vivo* electrophysiology.

Behavioral tasks. We used a classical conditioning paradigm in which odor cues predict specific outcomes. An odor was chosen pseudo-randomly (out of two or three odors) and was delivered for 1.5 second after which an outcome (1.5 µL of artificially sweetened water or no reward) was delivered via the waterspout. An inter-trial interval (ITI) separated each trial. The duration of the ITI followed a truncated exponential distribution with a mean of 8 seconds, minimum duration of

5 seconds, and maximum duration of 12 seconds. Each session consisted of 200 trials, which lasted about ~35 minutes on average.

In the stable task, the odor A (S)-(-)-limonene was always rewarded and odor B (1-heptanol) was always not rewarded. In the dynamic task, the reward contingency changed every session, with the reversal happening at 101th trial (middle point). In the hybrid task (for *in vivo* electrophysiology), we used 3 odors, two of which were stable (odor A and B) and one was dynamic (odor C; 1-hexanol). The outcome of the 3rd dynamic odor changed every 60 trials (instead of 100 trials), and the whole session lasted for 240 trials (total of 4 blocks). The stable cues (odor A and B) were presented 25% of the time and dynamic cue (odor C) were presented 50% of the time to counter-balance the ratio of stable and dynamic cues.

Odors. We used (S)-(-)-limonene (odor A), 1-heptanol (odor B), and 1-hexanol (odor C). For stable or dynamic task, we used odor A and B. For the hybrid task, we used odor A, B and C. All odors were diluted in mineral oil at 10%.

Behavioral training. Mice were first handled while undergoing water deprivation. This lasted for at least a week until mice reached around 85% of their baseline weight. We confirmed that mice were comfortable licking to a syringe that delivered water while handling, indicating a reduction in overall stress and a willingness to seek water. After handling, mice were habituated on the rig, for at least 3 days. We head-fixed the mice and immediately dispensed water reward. The 1st habituation session lasted for 15 minutes, and the subsequent habitation session lasted for 35 minutes (similar time as the actual length of a behavioral session). We confirmed on the third session that mice were comfortably licking to the spout. If mice never licked to the spout on the 3rd day, we continued the habitation until mice became comfortable licking.

For training in the stable task, we trained mice for at least 5 days. Mice were able to perform well on the first day and steadily improved (**Extended Data Fig. 1a, c**). Value memory for stable task was tested after at least 5 days of training. For training on the dynamic task, mice were first trained on the stable task for at least 3 days and then trained on the stable task for another 3 days with reversed contingency (odor A=no reward, odor B=reward). After this, mice were then trained on the dynamic task where reversal happened every session at trial 101, for at least 12 days. We made sure that every dynamic task session started with the contingency that the mouse had last seen from the previous session, to preserve the continuity of reward contingency across days.

For training in the hybrid task, mice were initially trained on the stable task for 5 days. Mice were then trained on the hybrid task by introducing the third cue (odor C). Initially, each cue was presented 33% of the time, and there were two blocks (one reversal) for the dynamic cue. After 8 days of training, we transitioned to the final version of the hybrid task with cue presentation ratio of 25/25/50% for cue A/B/C respectively, and with 4 blocks. After about 10 sessions in this final version of the hybrid task, Neuropixels recording was performed in the following sessions.

Testing value memory. We tested the stability of the value memory in two ways. In the first way, after mice were fully trained on the stable or dynamic task, we introduced break between sessions (1, 2, 4 or 8 days for stable task; 1 day for dynamic task). The break duration was randomized across mice. In the second way, after mice were fully trained on the stable or dynamic task, we increased the ITI length (duration=0.625, 1.25, 2.5, 5 min) on trial 50th and 150th. To quantify the mice's ability to remember previously learned values, we computed the discrimination index as follows:

$$discrimination\ index = \frac{CS^{+}lick(1st\ trial) - CS^{-}lick(1st\ trial)}{mean\ CS^{+}lick(last\ 5\ trials) - mean\ CS^{-}lick(last\ 5\ trials)}$$

A similar metric was used for dopamine photometry signal (Extended Data Fig. 1h) or for RNNs value (Main Fig. 2i).

Testing value inference. To test if mice could use structural knowledge of the task to infer a change in value based on value update for the opposite cue, we first trained mice fully on the dynamic task (anticorrelated). Next, at the reversal point, we presented one cue type for 5, 10 or 20 trials, followed by the other cue (probe trial). We quantified the change in the value of the probe cue relative to level before reversal. Any change in the value of the probe cue would indicate an inferred value update based on structural knowledge of the task (without direct experience). For mice, we computed the inference change in value by computing the change in CS licks (ΔCS licks) or in dopamine (DA) signal (ΔDA (norm. dF/F)). For RNNs, we quantified the change in inferred value (Δinferred value). For both mice and RNNs, the change was computed by subtracting the pre-reversal baseline level, which was computed by taking the mean value (CS licks, dopamine signal RNN value readout) over the last 5 trials before reversal.

We conducted additional experiments to test if mice could learn distinct correlation structures. Two different cohorts of mice were trained as described before but instead of the dynamic task having anti-correlated structure, the values of the odors were either positively correlated or independent. In the positively correlated structure, each session consisted of three 70-trial blocks: both odors rewarded, both unrewarded, then both rewarded again. In the independent structure, one odor was constantly rewarded whereas another odor was rewarded and then not rewarded, and vice versa in the next session. In order to test inference, one cue was presented 5-20 times before presenting the other cue. For the independent structure, the stable value odor was always the cue being tested for inference.

Fiber photometry. We used a commercially available bundle fiber photometry system (BFMC, Doric) to record photometry signals from multiple animals simultaneously. A low-autofluorescence Branching Bundle Patchcord (400μm, 0.57 NA) was connected to the photometry system. Each end of the fiber went to a single behavioral rig, allowing us to independent track photometry signals from up to 3 animals simultaneously. A blue excitation LED (470 nm, 10 μW power at tip of the 400um patchcord ferrule) was used to collect GRAB^{DA3m} signal. A purple excitation LED (415 nm, 9 μW power) was used to collect control signal for correcting movement artifact. The following parameters were used for imaging in the Doric Neuroscience Studio V6 software: power=10%, framerate=30Hz, exposure=0.012 seconds, gain=9.9dB.

KN-93 infusion experiment. For inhibiting synaptic plasticity in BLA, we used KN-93, a CaMKII inhibitor known to disrupt synaptic plasticity[40,53]. Once mice were fully habituated to the rig (for testing in stable task) or fully trained on the dynamic task (for testing in dynamic task), the day before the infusion, two small craniotomies were made above BLA on each hemisphere, and sealed with silicone elastomers (Kwik-Cast, World Precision Instruments). On the day of infusion, mice were lightly anaesthetized with isofluorane (1%). KN-93 (water soluble KN-93 dissolved in saline at 100 μM concentration) (422711-1MG, Millipore Sigma) was injected bilaterally into BLA (volume: 300 nl per site). Mice were left to recover for an hour before behavioral testing. To verify the injection site, KN-93 was mixed with a lipophilic tracer DiI (less than 5% of total volume) (V22889, ThermoFisher Scientific). Mice were perfused (see Histology) and the slices imaged under a wide-field microscope (see Extended Data Fig. 3). For saline infusion, the same procedure was performed with saline+DiI.

***In vivo* electrophysiology.** We performed acute Neuropixels (1.0, single shank) recording in mice fully trained the hybrid task. A total of 6 recording sessions was performed per mouse, 3 sessions per hemisphere. All recordings were performed in SpikeGLX software (https://github.com/billkarsh/SpikeGLX), with a sampling rate of 30 kHz, local field potential gain of 250 and action potential gain of 500, and we analyzed only the action potential channel (which was high-pass filtered in hardware with a cut-off frequency of 300 Hz). Behavioral and neural recordings were synchronized using a transistor–transistor logic (TTL) pulse sent from the Bpod to the PXIe acquisition module SMA input at the start of every trial.

The day before the first recording session, a small craniotomy was made bilaterally above BLA. Slicone elastomers (Kwik-Cast) was used to cover the craniotomies. On recording sessions, the silicone gel was gently removed to expose the brain. A Neuropixels probe was soaked into either DiO, DiI or DiD solutions (Vybran Multicolor Cell-Labeling Kit, ThermoFisher Scientific) for tracking the probe location, then mounted on a vertical manipulator (KDC101 and Z825B, Thorlabs). The probe was slowly lowered until it touched the brain surface. Saline was applied around the craniotomy to prevent drying and helping the insertion through the dura. The probe was slowly inserted into the dura at 0.05 mm/s. Once the probe was confirmed to have penetrated the brain, the speed was lowered to 0.01 mm/s. The probe was lowered to 5.3 mm below the brain surface to record in the BLA and regions around it. Once the probe was fully lowered, we waited for 10 minutes for the probe to fully settle inside the brain. Once the recording session was over, the probe was slowly retracted at 0.01 mm/s. Silicone gel was applied on the craniotomy.

Optogenetic experiments. We performed optogenetic experiments to test the role of BLA in stable and dynamic task. Optical fibers (400µm, 0.57 NA, Doric) were implanted bilaterally above BLA (coordinates: -1.3/3.3/4.0) in emx1-Cre × gtACR1 mice. Mice were allowed to recover for at least 2 weeks before starting the handling procedure. Mice were first trained on the stable task for 5 sessions, and then underwent 2 stimulation sessions. The same mice underwent 2 weeks of training on the dynamic task and then underwent 2 stimulation sessions. 470nm LED was used to inhibit BLA neurons. For testing the role of BLA during cue period, LED was on for a duration of 3 seconds aligned to odor onset, and randomly in 15% of all trials. for testing the role of BLA during ITI period, LED was on for a duration of 3 seconds, 0.5 seconds prior to the odor onset of the upcoming trial. We performed 2 stimulation sessions per mouse. We made sure to interleave a non-stimulation session to avoid chronic effect from repeated stimulation of BLA neurons during the dynamic task. The effect of stimulation was quantified by comparing the effect of stimulation on CS+ licks and CS- licks (Fig. 3d, Fig. 5h) or by quantifying the difference between CS+ licks and CS- licks (Fig. 5i).

RNN modelling.

Recurrent neural network implementation. We implemented RNNs as previously described[34,35]. Briefly, we trained recurrent neural networks composed of GRU units (N=20) to estimate value at each timestep. The hidden unit activity is given by the following equation:

$$z_t \ = \ f_\phi(o_t, z_{t-1})$$

z_t is hidden units' activity at time t

ϕ is the parameter vector of the RNN

o_t is the observation produced by environment at time t

z_{t-1} is the hidden units' activity at time t-1

o_t is a one-hot vector defined as follows:

$$\mathbf{o_t} = [odorA, odorB, outcome], where\ \ odorA \in \{0,1\}, odorB \in \{0,1\}, outcomeB \in \{0,1\}$$

Thus, at each timestep, the RNN had access to a vector indicating whether odor A was present, odor B was present, and whether reward was delivered or not. Each trial began with an intertrial interval (ITI) of duration that followed a geometrical distribution with parameter p=0.8. After the ITI, a cue was presented ($\mathbf{o_t} = [1,0,0]\ or\ \mathbf{o_t} = [0,1,0]$ for odor A or B respectively) followed by the outcome ($\mathbf{o_t} = [0,0,1]\ or\ [0,0,0]$) for reward or no reward respectively).$V_t = w^\top z_t + w_0$ for $z_t, w \in \mathbb{R}^H$ (where H=20 is the number of hidden units), $w_0, V_t \in \mathbb{R}$. The full parameter vector $\theta = [\boldsymbol{\phi}, \boldsymbol{w}, w_0]$ was learned using TD learning. This involved backpropagating the gradient of the squared error loss $\delta_t^2 = (r_t + \gamma V_{t+1} - V_t)^2$ with respect to V_t. The discount factor γ was set to γ=0.2.

Task implementation. To implement the stable task, we used 200 trials where each odor A or B was chosen at random, with odor A being rewarded and odor B not being rewarded. For the dynamic task, reward contingencies flipped every 50 trials (block length = 50 trials). The dynamic task lasted for 18 blocks.

Training. We first initialized the network to PyTorch's default. We used a truncated backpropagation through time (T-BPTT) learning rule[54,55] where at each timestep, the network used recent inputs (defined by the window size W) to compute the gradient and update the weights of the network. We explored the effect of varying window size W (see Extended Data Fig. 2). For the majority of the analyses we fixed the window size to W=720 (~100 trials), which was roughly equal to the length of two blocks in the dynamic task. Given that the network is continuously learning in this setting, it is possible that the learning gets stuck in non-optimal local minima. To avoid such networks that might have sub-optimally learned the task, we first computed mean loss (squared reward prediction error) for the last 20 trials of the last four blocks in the dynamic task for a range of W (Extended Data Fig. 1a-b). Most networks had a loss less than 0.0005. Thus, we excluded networks whose loss exceeded 0.0005. This was to ensure that the final RNNs could solve the task, regardless of the mechanism being used. Learning rate was set to 0.0005, and we used the Adam optimizer with the AMSGrad variant (amsgrad=True in PyTorch), which enforces a non-decreasing second-moment estimate for more stable updates.

Plasticity manipulation. We manipulated the plasticity in RNNs to test the role of plasticity in updating value in either stable or dynamic task. The learning rate was set to zero before training on the stable task to test the role of plasticity in the stable task. For the dynamic task, RNNs were first fully trained on the dynamic task, while excluding RNNs that never converged (see **Training**). We then set the learning rate to zero to test the role of plasticity in updating value in the dynamic task. To quantify the effect of plasticity manipulation, we computed the area under the receiver operating characteristic (AUROC) between value readout of CS+ cue and CS- cue.

PCA analysis. We applied principal component analysis (PCA) to better understand the neural state-space trajectories during stable and dynamic task. PCA was applied to the hidden units' activity over the entire period of training. Example state-space trajectories were plotted in the PC1-PC2 and value readout space. Note that the value readout space is not strictly a function of hidden units' activity, but a combination of hidden units' activity and the readout weights.

Context axis analysis. We defined the context axis as the axis that could best discriminate the context (block type) in the hidden unit activity space in the dynamic task. We first took the ITI activity of the last 4 blocks of the entire 18 blocks training of the dynamic task. The activity was then segregated into two contexts, and then a Fisher linear discriminant analysis was used to compute the line that best separate context. The resulting axis was defined as context axis. Activity during the ITI was projected onto the context axis for further analysis ($\text{context}^{\text{proj}}$).

To understand how the context information was being used, we computed the Spearman rank correlation coefficient between the $\text{context}^{\text{proj}}$ and the value difference between the last two cues (value readout of last cue A – value readout of last cue B). The correlation was computed using the first 4 blocks for naïve RNNs or using the last 4 blocks for expert RNNs.

Long ITI effect. To understand whether long ITI could have differential effects on the value memory in RNNs, we simulated the effect of long ITI by increasing the length of ITI to 5, 10, 20 or 40 in either stable or dynamic task. Discrimination index was computed similarly for RNNs (see **Testing value memory**).

To understand the effect of the long ITI, we projected the activity during the long ITI onto the context axis. We computed the drift by computing the distance moved along the context axis during the long ITI. We then computed the correlation coefficient between the value differential after the long ITI and the drift along the context axis. Value differential was corrected to be positive at the beginning of the long ITI. $\text{Context}^{\text{proj}}$ was also corrected to be positive if drift was moving away from the initial point towards the other fixed point.

Analysis

Photometry preprocessing. dF/F was first calculated by computing F_0, and using the formula dF/F=(F_{raw} – F_0)/F_0. F_0 was defined as the 10th percentile of F_{raw} in a rolling window of 30 s. dF/F traces were upsampled from 20 Hz to 1000 Hz through linear interpolation (MATLAB function interp1) and then smoothed with a Gaussian filter (SD = 50 ms). We then normalized dF/F by z-scoring (MATLAB zscore function).

Spike sorting. Neuropixels recording data were spiked sorted offline with Kliosort4 (https://github.com/MouseLand/Kilosort?tab=readme-ov-file) with default parameters, followed by manual curation of individual units using Phy (https://github. com/cortex-lab/phy).

Brain registration for *in vivo* electrophysiology. To label individual units with their corresponding location in the brain, we registered the histology to the Allen Mouse Brain Atlas. We first used AP_histology to register each histology slices with the tracer to the Allen Mouse Brain Atlas (https://github.com/petersaj/AP_histology/tree/master). Coordinates for each probe track was then converted in the relevant format to be read out by the IBL Ephys Atlas alignment tool (https://github.com/int-brain-lab/iblapps/tree/master/atlaselectrophysiology). We obtained a brain location for each recorded single unit, which was used for brain region specific analysis.

Histology. Mice were deeply anesthetized with an overdose of ketamine/medetomidine, exsanguinated with 0.9% phosphate buffered saline (PBS), and transcardially perfused with cold 4% paraformaldehyde (PFA) in PBS. The brain was extracted from the skull and stored in 4% PFA for 24-48 hours at 4°C, after which it was rinsed with PBS, stored in PBS, and cut into 100 μm sections on a vibratome (VT1000S, Leica). Sections mounted on slides and then imaged using a slide scanner (Zeiss Axioscan 7).

Neuropixels recording analysis

Value coding cells. We defined a cell as stable value coding if it passed the following criteria: significantly different firing rate between odor A and odor B during the 1.5s odor on period. We tested for significance using student's *t*-test (MATLAB ttest2 function) at α=0.01, for different blocks combination (block1+2, block 3+4, block1+4, block2+3, block1+2+3+4). A cell had to pass all 5 tests to qualify as a stable value coding cell. This was to ensure that the cell was firing consistently across blocks. For dynamic value coding, we similarly defined dynamic value coding cell if it passed the following criteria: significantly different firing rate between odor C during rewarding block and odor C during non-rewarding block. We tested for significance using student's *t*-test at α=0.01for different combinations of blocks (block 1 vs 2, block3 vs 4, block1+3 vs 2+4). A cell had to pass all 3 tests to qualify as a stable value coding cell. This was to ensure that the dynamic value was consistently maintained throughout the session and not just in specific blocks. Lastly, a stable and dynamic value coding cell was defined as a single unit that was both stable value coding and dynamic coding following the criteria mentioned above. When quantifying the percentage of cells encoding value, we excluded cells whose baseline firing rate (total spikes/session duration) was below 0.2 Hz.

Context coding cells. We defined a cell as context coding if it passed the following criteria: significantly different firing rate during the ITI (4 seconds before cue onset till cue onset) between rewarding block and non-rewarding block. We tested for significance using a student's *t*-test at α=0.01, for different blocks combination (block 1 vs 2, block3 vs 4, block1+3 vs 2+4). A cell had to pass all 3 tests to qualify as a context coding cell.

Correlation between CS licks and context coding. To quantify the relationship between the context coding during the ITI and value for the dynamic cue, we computed the Spearman correlation between the ITI firing rate during the ITI (last 4 seconds before odor onset) and the number of licks during odor C delivery (1.5 seconds duration). We combined trials from block2 and block4 or block 1 and block 3 and computed the correlation for the two sets of trials, and then computed the mean correlation across the block types. This was to avoid introducing spurious correlation given that context coding was initially defined by being able to differentiate block 1+3 vs block 2+4, and given that mice's licks to odor C also differentiated block type. For control, we shuffled the CS licks for odor C and ITI firing rate pairing.

Context update analysis. In the hybrid task, average reward rate and context coding are confounded. To determine if context coding reflected block identity based on the outcome of odor C only, we quantified the amount of context coding update for each cue type. If context coding neurons reflect reward rate, then one would expect context to be updated solely based on outcome regardless of cue type. However, if context coding truly reflects odor C specific context, then one would expect context coding to be only updated after odor C. We computed the mean context update at the beginning of each block (Δcontext coding) by first taking the change in ITI firing rate (4 seconds before odor onset till odor onset) after each trial type. We focused on the first six trials for each cue type because context update occurred mostly at the beginning of each block. We restricted our analysis for block 2 and block 4. Thus, for each context coding unit, we obtained a mean context update for each cue type, which was taken from a total of 12 trials for each cue type. Predictions for the reward rate model vs pure context coding model is shown in Extended Data Fig. 5.

Figure + legends

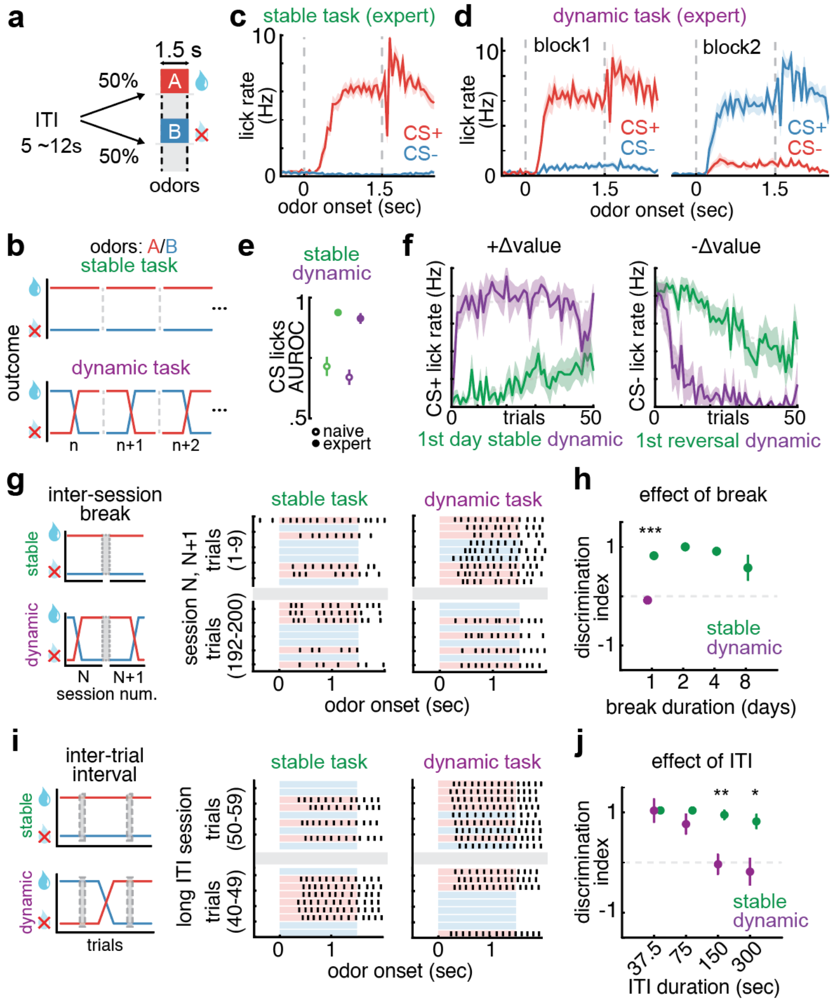


Fig. 1 | Environmental stability sets distinct timescales for value update and forgetting. a, Trial structure. Following an ITI (5~12 s), an odor cue (1.5s) was presented randomly (odor A or B), followed by an outcome (water reward or no reward). **b,** Task type. In the stable task (top), the outcome was fixed (odor A-reward, odor B-no reward). In the dynamic task (bottom), the outcome flipped once every session (2 blocks). Each session started with the same reward contingency as the last block from the previous session. **c,** Expert performance of mice trained on the stable task (training duration=3days). Lick rate (Hz) for rewarded odor CS+ (blue) and unrewarded odor CS- (red) is shown aligned to odor onset (n=10 mice). **d,** Similar quantification as in **c** for expert mice on the dynamic task, for block 1 (left) and block 2 (right) (n=10 mice). **e,** AUROC between CS+ and CS- in stable (green) and dynamic (purple) task for naïve (empty circle) and expert mice (filled circle) (n=10 mice; ***, *P*<0.001, two-tailed *t*-test). **f,** Learning curve for updating value (+Δvalue:

positive update, –Δvalue: negative update) in stable/1st reversal and dynamic task. *left*, CS+ lick rate for naïve mice first exposed to the stable task, aligned to 1st day (green) and expert mice on the dynamic task, aligned to reversal point (0 trial=reversal point) (purple). *right*, similar quantification for CS- lick rate: mice experiencing reversal for the first time (green) aligned to reversal point and expert mice on the dynamic task decreasing CS- lick rate aligned to reversal point (stable: n=6 mice; dynamic: n=5 mice). **g,** Testing the timescale of value forgetting in dynamic vs stable task with inter-session break. *left,* schematic showing the break (grey box) separating session N from session N+1. *right*, lick raster plot for an example session aligned to odor onset for stable and dynamic task. Grey box is the break (24 hours) between session N and session N+1. Each row indicates a trial with the colored box indicating the odor on period (red, CS+; blue, CS-). **h,** Quantification forgetting for different inter-session break duration (1, 2, 4, or 8 days break). Discrimination index represents how well mice could discriminate CS+ and CS- on the first trial for each CS on session N+1 (normalized by previous session performance; see Methods) for stable (green) or dynamic (purple) task. (stable: n= 4 mice; dynamic: n= 7 mice; ****,* *P*<0.001, two-tailed *t*-test). **i,** Testing the timescale of value forgetting in dynamic vs stable tasks with inter-trial break. left, schematic showing two breaks within a session (grey boxes). *right*, lick raster plot for an example session similar to **g**. **j,** Similar quantification as in **h** for inter-trial breaks (37.5, 75, 150 or 300 sec) (stable: n=6 mice; dynamic: n=7 mice; **,* *P*<0.05; ***,* *P*<0.01, two-tailed *t*-test). All data shown are mean ± s.e.m.

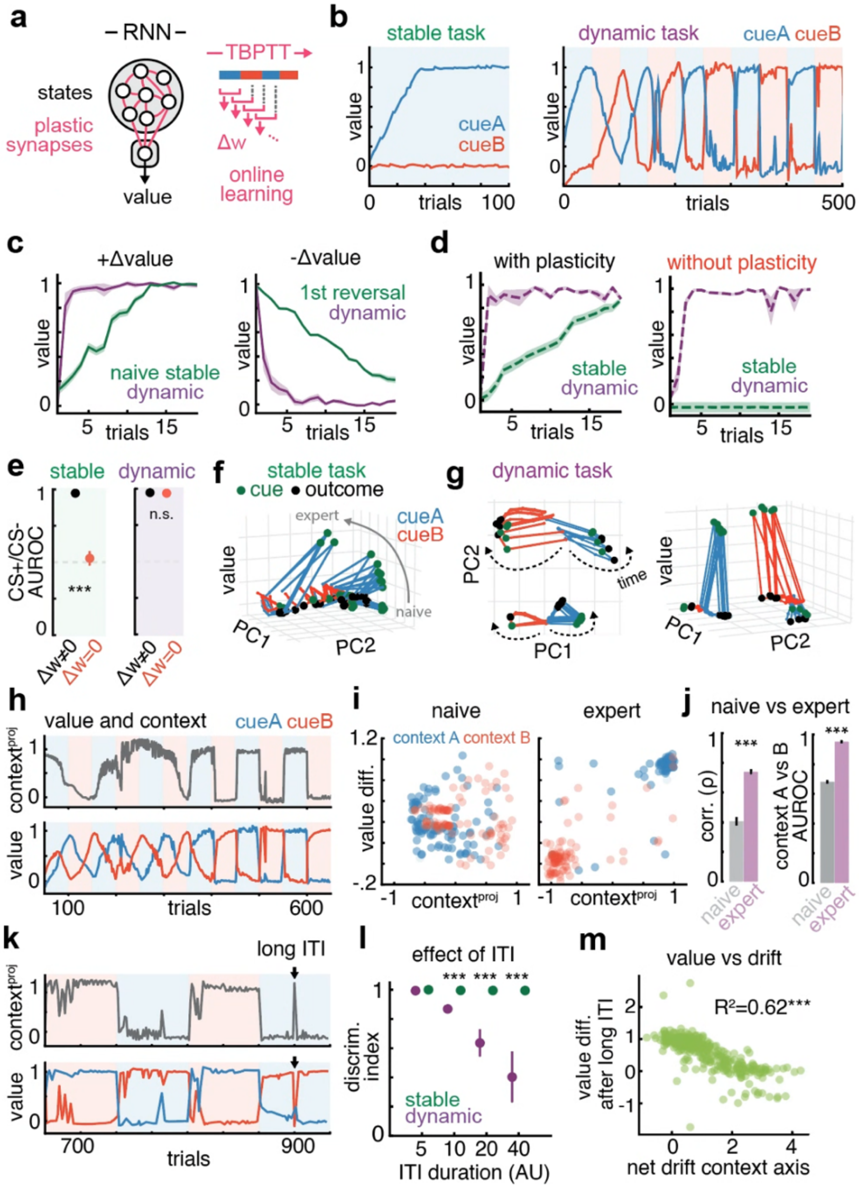


Fig. 2 | RNNs with online weight update recapitulates mouse behavior. a, *left*, Schematic showing the RNN with plastic synapses (pink). States are represented by the hidden units' activity, and value is readout using RNN representation (see Methods). *right*, Schematic showing the online learning rule using TBPTT. At each timestep, the RNN updates its weight based on the past experience using a sliding window (see Methods). **b,** *left,* Example RNN learning the stable task. Blue and red traces re the RNN value readout for rewarded cue (CS+) and unrewarded cue (CS-) respectively. *right,* Example RNN learning the dynamic task (blue, cue A; red, cue B). Background color denotes the cue type that is being rewarded in that block. **c,** *left,* Example RNNs (n=10) showing positive value update (+Δvalue) in the stable task (green) and in the expert level dynamic task (purple) For stable task, value starts from naïve to first exposure to the task. For dynamic task, value starts from the start of reversal for the cue that was previously unrewarded. *right,* similar quantification for (–Δvalue) in the first reversal (green) and in expert level dynamic task (purple). **d,** Example RNNs (n=10) showing the effect of freezing the weight in the stable (green) or dynamic task (purple) for positive value update (black=weight update intact; red=weight frozen). **e,** Quantification of the AUROC for CS+ and CS- value readout in the stable (left) and dynamic (task) when weight update was intact (black) or frozen (red). Weight freezing had a significant effect on updating value in the stable task (****, P*<0.001, two-tailed *t*-test) but no effect in the dynamic task (*P*=0.31, two-tailed *t*-test). **f,** Example neural trajectories in PC1, PC2 and value space during stable task for the first 30 trials. Green dot indicates the time of cue and black dot indicates the time of outcome delivery. Cue A (reward cue) is shown in blue and cue B (non-rewarded cue) is shown in red. **g,** Similar plot as in **f** for an expert RNN trained on the dynamic task. *left*, example neural trajectories are shown for both block. PC1 separate cue type whereas PC2 separates block type. *right*, same neural trajectories plotted in PC1, PC2 and value space. **h,** An Example RNN with the hidden units' activity during the ITI projected onto the context axis (context^{proj}) (*top*, see Methods) and the value readout (*bottom*) for cue A (blue) and cue B (red). Colored background represents which cue is being rewarded in that block. **i,** Quantification of the correlation between the value differential (value diff.) and the context projection (context^{proj}) for naïve (*left*) and expert (*right*) RNN (n=1). Value differential which was defined as the value difference between the last CS+ and CS- (see Methods). Each dot represents a single trial (red/blue=types of blocks). **j,** *left,* Quantification of Spearman's rank correlation coefficient (ρ) between value diff. and context^{proj} in naïve (grey) vs expert (purple) RNNs (n=100). Context projection became more predictive of value differential in expert compared to naïve RNNs (****P*<0.001, two-tailed *t*-test). *right,* AUROC between block types (cue A-reward block vs cue b-reward) in naïve (grey) vs expert (purple) RNNs (n=100). Context projection became more separable in naïve vs expert mice (****, P*<0.001, two-tailed *t*-test). **k,** An example expert RNN on the dynamic task, quantified similar to **h,** when long ITI is introduced (black arrow). **l,** Quantification of the effect of introducing long ITI (discrimination index) as a function of the length of the ITI (5-40AU) for stable (green) and dynamic task (purple). Long ITI trials decreased the discrimination index for dynamic but not stable task (***, *P*<0.001, two-tailed *t*-test). **m,** Relationship between value difference (value diff.) between CS+ and CS- after long ITI and the drift caused by the long ITI along the context axis (see Methods). The amount of drift on the context axis predicted the change in value readout between CS+ and CS- (R^2=0.62, two-tailed Pearson's correlation coefficient test, ***, P<0.001). All data shown are mean ± s.e.m.

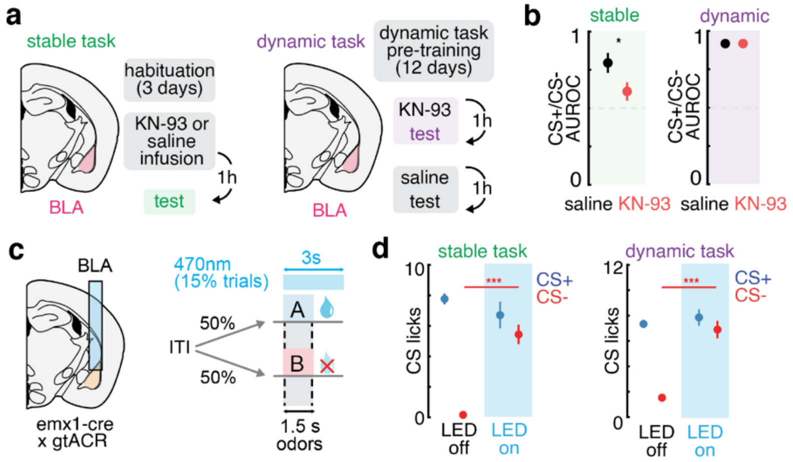


Fig. 3 | Dissociable roles of BLA plasticity vs activity. a, Schematic showing experimental flow for testing the role of BLA plasticity in stable vs dynamic task. *left,* Mice were first habituated to the rig for 3 days, after which they underwent infusion of either KN-93 or saline, targeting BLA (pink). Mice were tested on the stable task after 1h post-infusion. *right,* Mice were trained on the dynamic task for 12 days until they reached expert performance, after which they underwent infusion of KN-93 and saline in the BLA on different consecutive sessions (see Methods). **b,** Effect of KN-93 infusion on the performance in stable (left, green) or dynamic (right, purple) task. AUROC between CS+ and CS- licks are quantified for each experiment. (stable/saline, n=11 mice; stable/KN-93, n=8 mice; dynamic, n=7; *, P<0.05, two-tailed t-test). **c,** Schematic showing experimental flow for inhibiting BLA activity in the stable vs dynamic task. *left,* emx1-Cre × gtACR1 mice were implanted with bilateral fibers above BLA. *right*, stimulation light (470 nm LED) was on for 3 sec starting from odor onset on 15% of all trials. **d,** Quantification of the effect of inhibiting BLA activity in stable (left) and dynamic (right). CS+ (blue) and CS- (red) licks are shown for LED off (left) and LED on (right) (n=10 sessions form 5 mice; ***, P<0.001, two-tailed t-test). All data shown are mean ± s.e.m.

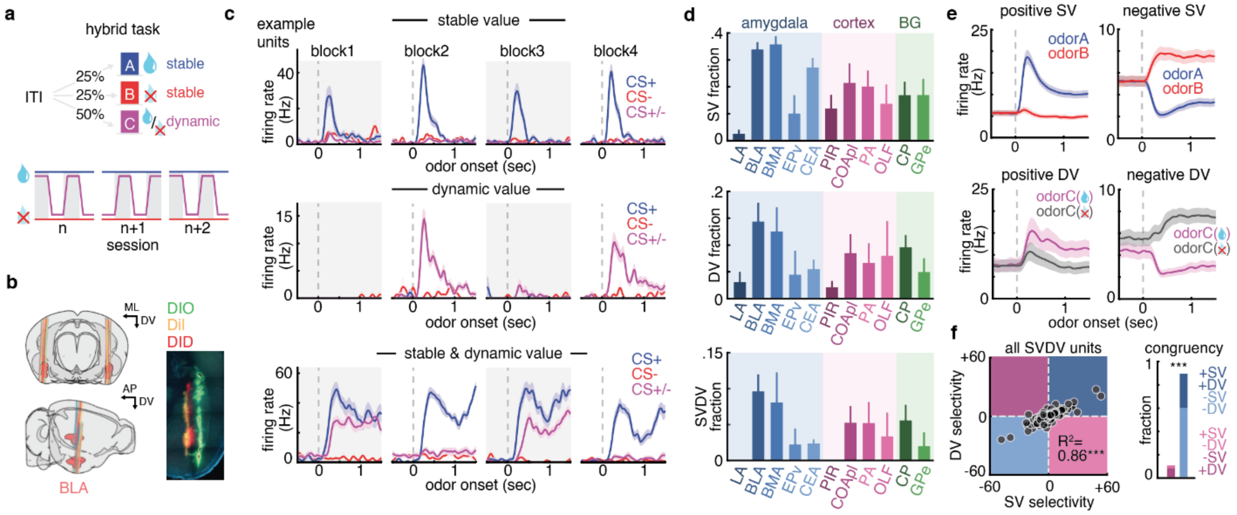


Fig. 4 | Stable and dynamic value coding in the BLA. a, Schematic showing trial structure and reward contingency across session. *top*, in the hybrid, task, three odors were presented, odor A or odor B with 25% probability and odor C with 50 % probability. Odor A and B had stable value (A-reward, B-no reward) whereas the value of odor C was dynamic. *bottom*, Each session consisted of four blocks where the value of odor C alternated between reward and no reward. Initial value of odor C was kept the same as the value of odor C in the last block of previous session to preserve continuity. **b,** Neuropixels probe trajectories targeting BLA. *left*, 3d rendering of the probe trajectories of recording sessions. BLA is shown in red. *right*, example histological slice showing probe trajectories. DiO (green), DiI (yellow) and DiD (red) were used probe tracking (see Methods). **c,** Example units coding for stable value (top), dynamic value (middle), or stable and dynamic value (bottom). Each graph shows the smoothed peri-stimulus time histogram (PSTH) averaged across each block (grey, odor C → reward; white, odor C → no reward). Each trace shows firing rate (spikes s^{-1}) for CS+ odor (blue), CS- odor (red) or dynamic CS+/- (purple). **d,** *top,* Fraction of units encoding stable value (SV, top), dynamic value (DV, middle), or stable and dynamic value (SVDV, bottom). **e,** *top,* Mean firing rate of all amygdala positive (left, n=780) or negative (right, n=1146) SV neurons, aligned to odor onset (blue/red, CS+/CS-). *Bottom,* Mean firing rate of all positive (left, n=344) or negative (right, n=510) DV neurons aligned to odor onset (purple/grey, CS+/CS-). **f,** *left,* Stable value selectivity (firing rate^{A} – firing rate^{B}) vs dynamic value selectivity (firing rate$^{\text{C-Reward}}$ – firing rate$^{\text{C-No reward}}$) for all DV neurons (R^2=0.62, two-tailed Pearson's correlation coefficient test; ***, P<0.001). *right,* Fraction of units categorized by the polarity of SV and DV selectivity (SV+DV+, dark blue; SV-DV-, light blue; SV+DV-, light pink; SV-DV+, dark pink). All data shown are mean ± s.e.m.

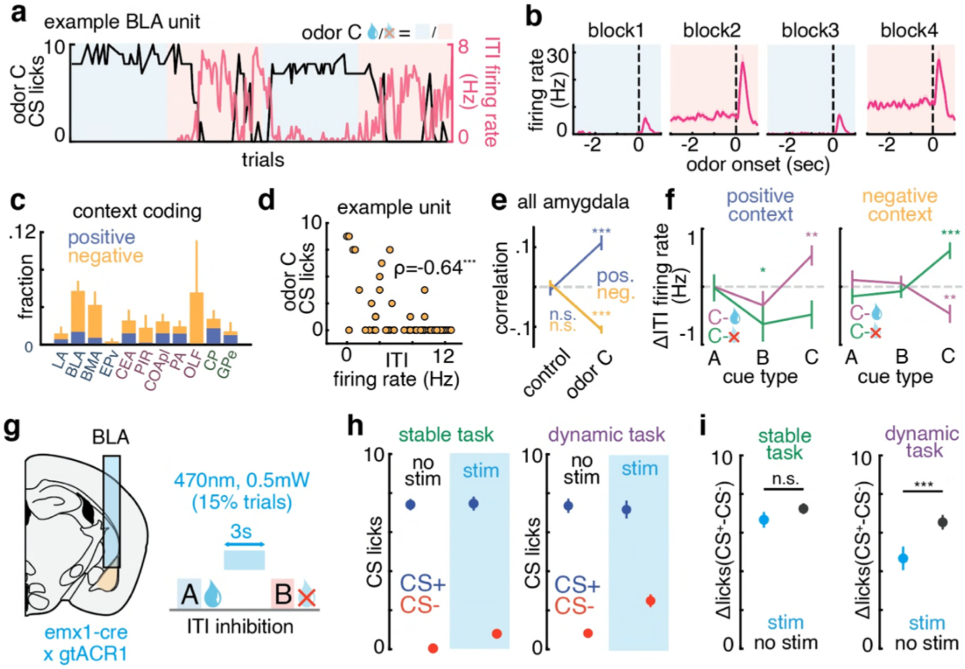


Fig. 5 | Context coding in the BLA. a, Example BLA unit that encodes context in the ITI. CS licks to odor (black line) is shown along the mean ITI firing rate during the 3 seconds before cue onset (pink line, see Methods). Background colored rectangles indicate the block type (light blue, rewarded odor C; light pink, not rewarded odor C). **b,** Same example BLA unit as in **a** showing mean firing rate during each block aligned to odor onset (background color indicates block type as in **a**). **c,** Fractions of units context coding units across all brain regions recorded. Positive/negative context (blue/yellow) was defined by the whether ITI firing rate was higher in the odor C reward block vs odor C no reward block (see Methods). **d,** Example negative context encoding BLA unit showing the correlation between CS licks to odor C and ITI firing rate before odor C onset. Data is from block 2+4 combined, showing correlation within same block type (see Methods) (ρ=-0.64, two-tailed Spearman's rank correlation test; ***, *P*<0.001). **e,** Quantification of the mean Spearman's rank correlation between odor C CS licks and ITI firing rate for all positive (pos., blue, n=59) and negative (neg., yellow, n=171) in the amygdala (***, *P*<0.001, two-tailed *t*-test). For control, we shuffled the CS licks-ITI firing rate pairing (see Methods). **f,** ITI firing update for all positive (left) and all negative (right) context coding amygdala units. At the beginning of each block, we quantified the change in ITI firing rate followed by each cue type (odor A, B or C) in either odor C reward block (purple) or odor C no reward block (green) (*, *P*<0.05, **, *P*<0.005; ***, *P*<0.001, two-tailed *t*-test). **g,** Schematic showing experimental flow for inhibiting BLA activity during the ITI period. *left,* emx1-Cre × gtACR1 mice were implanted with bilateral fibers above BLA. *right*, stimulation light (470 nm LED) was on for 3 sec during the ITI on 15% of all trials. **h,** Quantification of CS licks (blue: CS+; red: CS-) for no stim trials (black) or trials where stimulation was on preceding cue (blue area) for stable task (left, n=12 sessions) or dynamic task (right, n=10 sessions). **i,** Quantification of the difference in CS licks (Δlicks (CS^+-CS^-)) for stim (light blue) and no stim trials (black) in stable (left) or dynamic (right) task (n.s., $P>0.05$; ***, $P<0.001$, two-tailed *t*-test). All data shown are mean ± s.e.m.

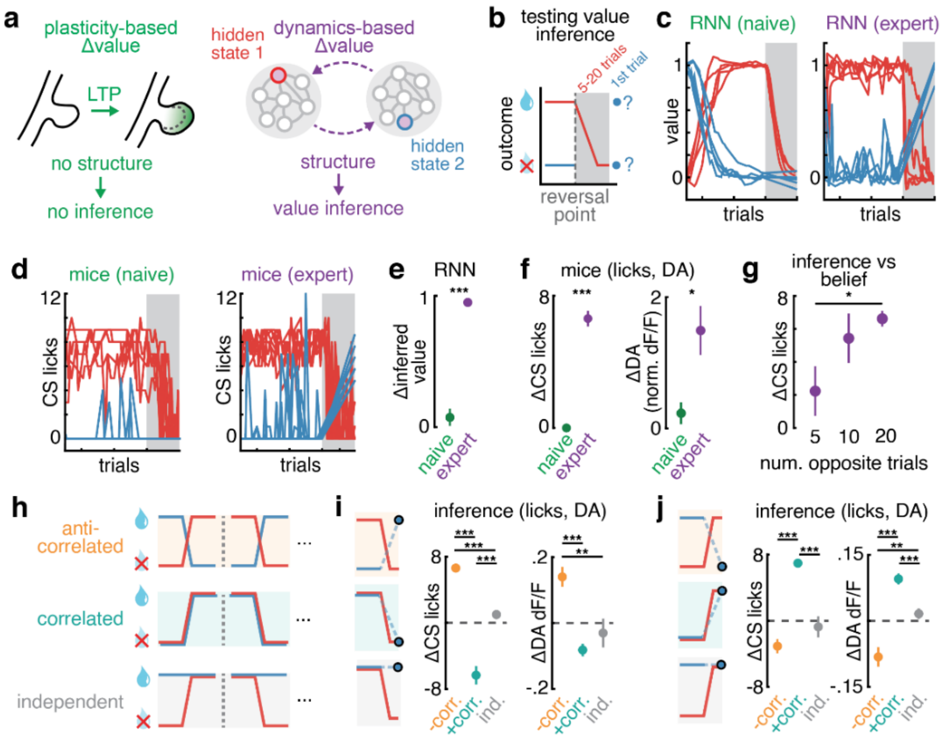
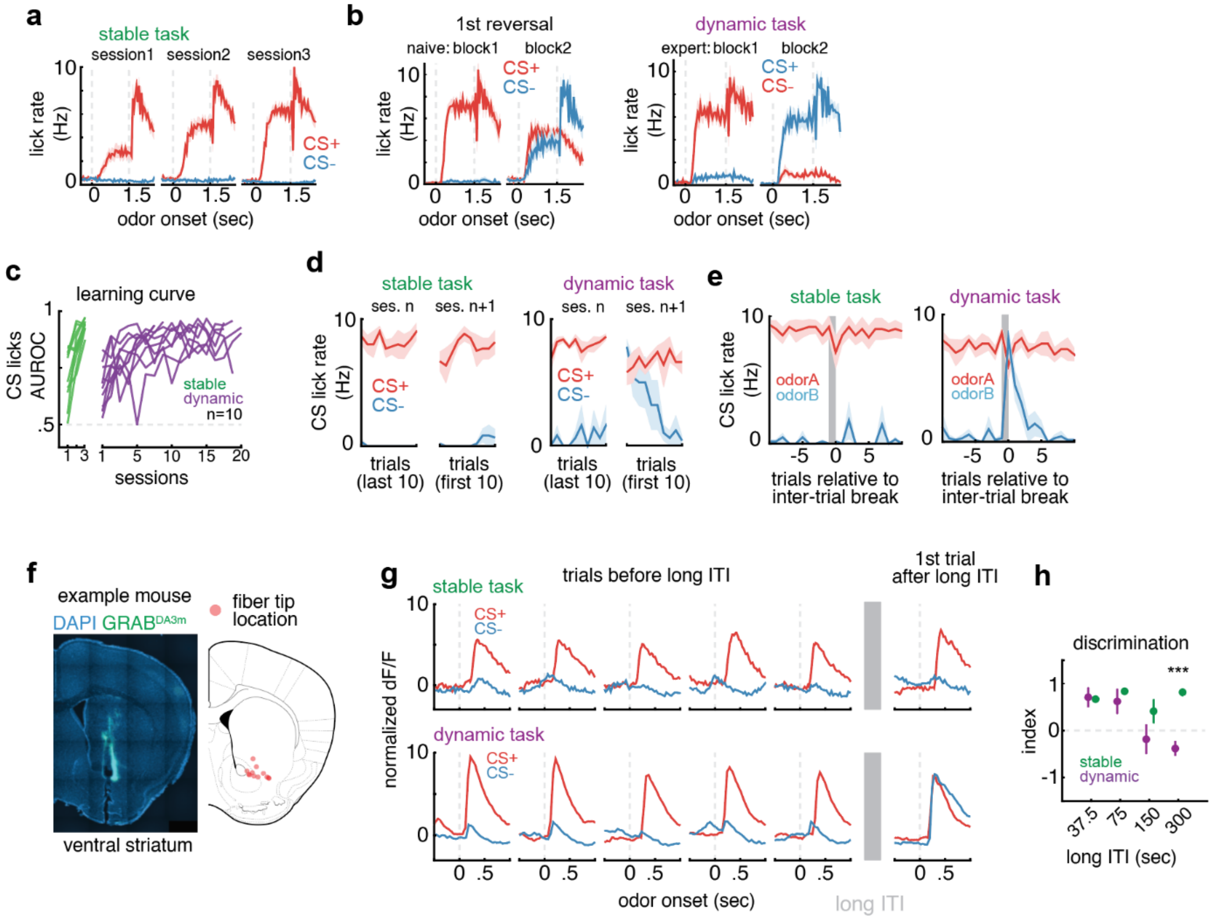
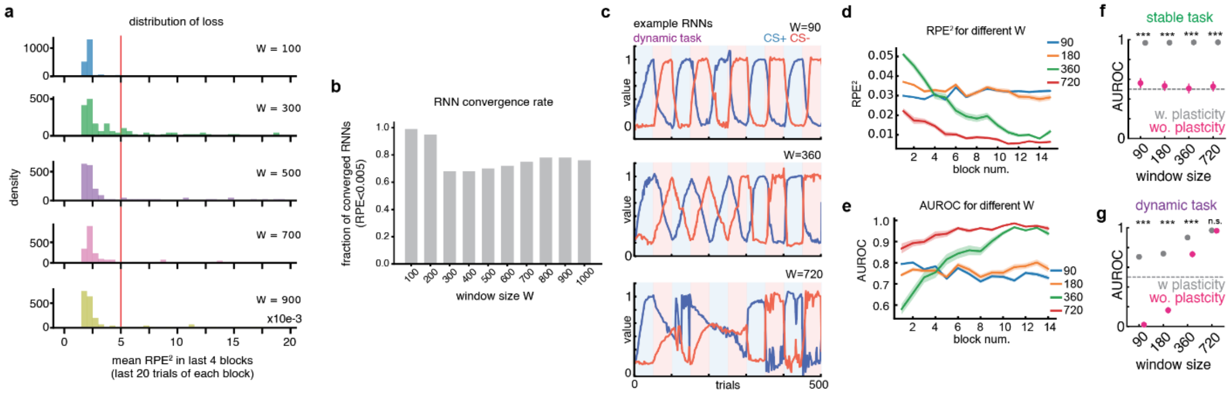


Fig. 6 | Recurrent dynamics enable structure-specific value inference. a, Schematic showing distinct predictions for value inference. Plasticity-based value update (left) has no learned structure, thus cannot infer value. Dynamics-based value update (right) learns the hidden states of the environment. The learned structure enables value inference. **b,** Task design to test value inference. Once mice were fully trained on the dynamic task, mice were tested around the reversal point by only presenting one cue type (red, 5-20 trials), after which the other cue was presented (blue). Value (e.g. CS licks) of the other cue could be inferred to be high using structural knowledge (value inference). **c,** Testing value inference in naïve (left) vs expert (right) RNNs trained on the dynamic task (n=4 example RNNs). **d,** Testing value inference in naïve (left) or expert (right) mice on the dynamic task (n=6 mice). **e,** Quantification of change inferred value in RNNs (Δinferred value) in naïve (green, n=100 RNNs) and expert (purple, n=100 RNNs) RNNs (***, P<0.001, two-tailed t-test). **f,** Quantification of inferred value in mice using CS licks (left, n=6) or dopamine photometry signal (right, n=6 mice) (*, P<0.05; ***, P<0.001, two-tailed t-test). **g,** Relationship between ΔCS licks and the number of trials of the opposite cue, indicating strength of belief that a change of state has occurred (n=6 mice) (*, P<0.05, two-tailed t-test). **h,** schematic showing three distinct correlations structures of odor values used for dynamic task (see Methods). **i,**

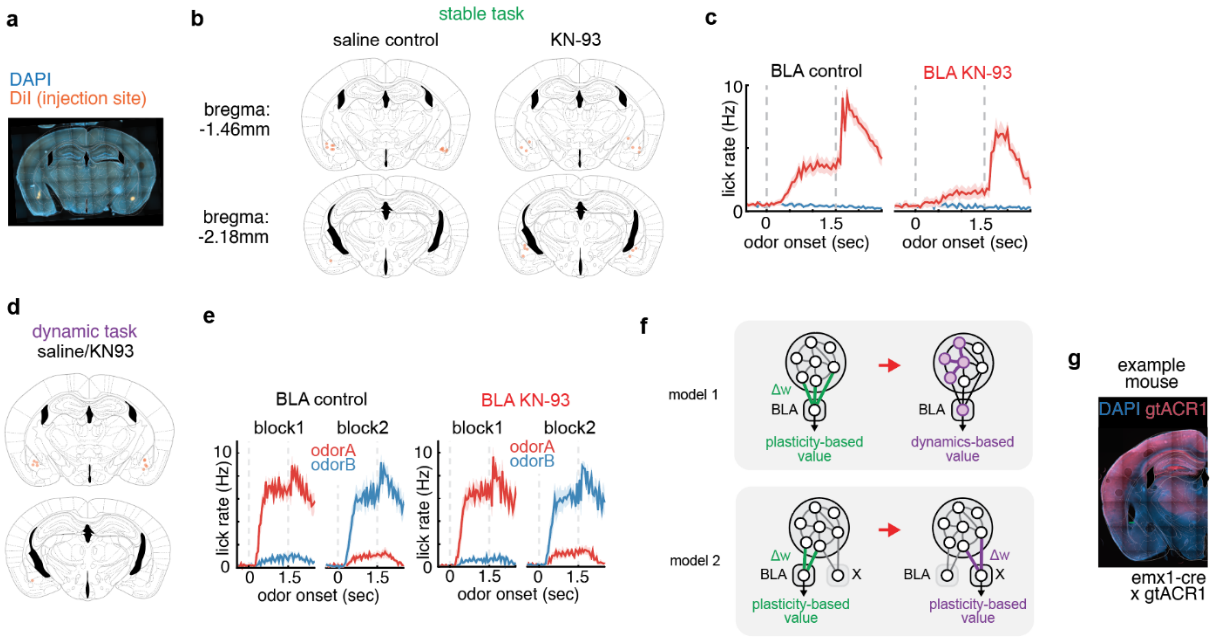
quantification of changed in inferred value for anti-correlated (-corr., orange), correlated (+corr., turquoise), or independent (ind., gray) structures, when the value of the opposite cue decreased. Change in inferred value is quantified using either licks during CS (ΔCS licks) or dopamine signals (ΔDA dF/F) (-corr.=12, +corr.=6, ind.=6). **j,** similar quantification as in **i** when the value of the opposite cue increased (*, *P*<0.05, **, *P*<0.005; ****,* *P*<0.001, two-tailed *t*-test). All data shown are mean ± s.e.m.



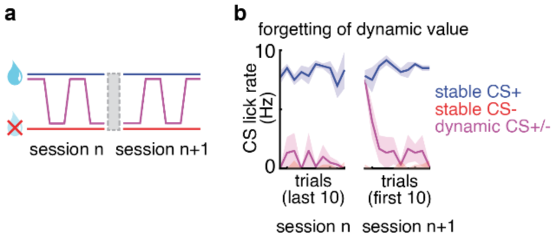
Extended Data Fig. 1 | Further quantification of behavioral training, forgetting after breaks, and dopamine signaling related to forgetting. a, Lick rate for session 1-3 for the stable task. Mice could learn to discriminate from session 1, with performance gradually improving over 3 days (n=10 mice) **b,** Comparing naïve vs expert performance on the dynamic task. *left*, 1st reversal session for block 1 and block 2. *right,* similar quantification for expert performance. Naïve mice initially had difficulty updating value and adapting to the new reward contingency on block 2 (n=10 mice) **c,** Quantification of the AUROC for CS+ and CS- for stable dynamic task shown as a function of session number (n=10 mice). **d,** Effect of inter-session break (1 day) for stable (left) or dynamic (right). CS lick rate is plotted for each odor (A, red; blue, B) as a function of trial number (last 10 trials or first 10 trials for session N or session N+1 respectively). Mice start licking to the CS- odor after 24 hours break in the dynamic task but gradually learn to suppress licking to the CS-. This effect is not present in the stable task (stable: n=4 mice; dynamic: n=7 mice). **e,** similar plot as in **d** for inter-trial break (300 sec). Grey box indicates the long break (stable: n=6 mice; dynamic: n=7 mice). **f,** *left,* Histological slice of am example mouse showing the expression of the dopamine sensor GRAB$^{\mathrm{DA3m}}$ (green) in the ventral striatum. *right,* fiber locations (red) aligned to the atlas for all mice (n=12 mice). **g,** Example sessions showing dopamine response to CS after long ITI (300 sec). *top*, Normalized dF/F for CS+ (red) and CS- (blue) aligned to odor onset for stable task. The plot shows the last five trials (column=trial) before the long ITI (grey box). After long ITI, the dopamine response to the CS+/CS- are similar to the level before. *bottom*, similar quantification for dynamic task. After long ITI, the dopamine response to the CS- increases to the level similar to the CS+ response. **h,** Quantification of the forgetting timescale as in **Fig. 1j** for CS dopamine response (mean normalized dF/F during odor period: see Methods). Discrimination index was lower in dynamic task compared to stable task for longer ITI duration (300 sec) (n=7 mice; ****, P*<0.001, two-tailed *t*-test).



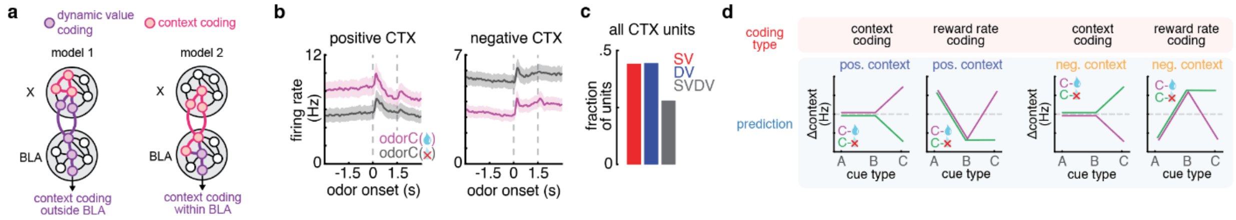
Extended Data Fig. 2 | RNN modelling with different hyperparameters. a, Histogram showing the distribution of mean loss (mean RPE^2 in the last 4 blocks, only computing the last 20 trials for each block) for different window size (W=100, 300, 500, 700, 900) during the dynamic task. Red line represents the threshold (0.005) for convergence (see Methods). For each window, we trained n=100 RNNs. **b,** Histogram showing the fraction of RNNs (mean RPE^2 <0.005) for different window size W. **c,** Value readout of CS+ (blue) and CS-(red) of example RNNs trained on the dynamic task with different window sizes (top: W=90, middle: W=360, bottom: W=720). RNNs with short W=90 do not display meta-reinforcement learning (speeding up of value update). RNNs with medium sized W=360 displays meta-reinforcement learning where the value update transitions from slow to fast process. RNNs with long W=720 displays meta-reinforcement learning but struggles to update value prior to meta-reinforcement learning due to the window encompassing multiple blocks. **d,** Quantification of the mean RPE^2 for each block in the dynamic task. Each color indicates a different window size (n=100 RNNs for each window). **e,** Quantification of the AUROC of value readout of CS+ and CS- for each block in the dynamic. Each color indicates a different window size (n=100 RNNs for each window). **f,** Effect of freezing weight update on the value update during the stable task for different window size. AUROC is plotted against window size for RNNs network with plasticity (w. plasticity: grey) and without plasticity (wo. plasticity: pink). Dotted line represents chance level (AUROC=0.5) (***, *P*<0.001, two-tailed *t*-test) **g,** Similar quantification as **f** but for the dynamic task (***, *P*<0.001, two-tailed *t*-test; n.s., not significant).



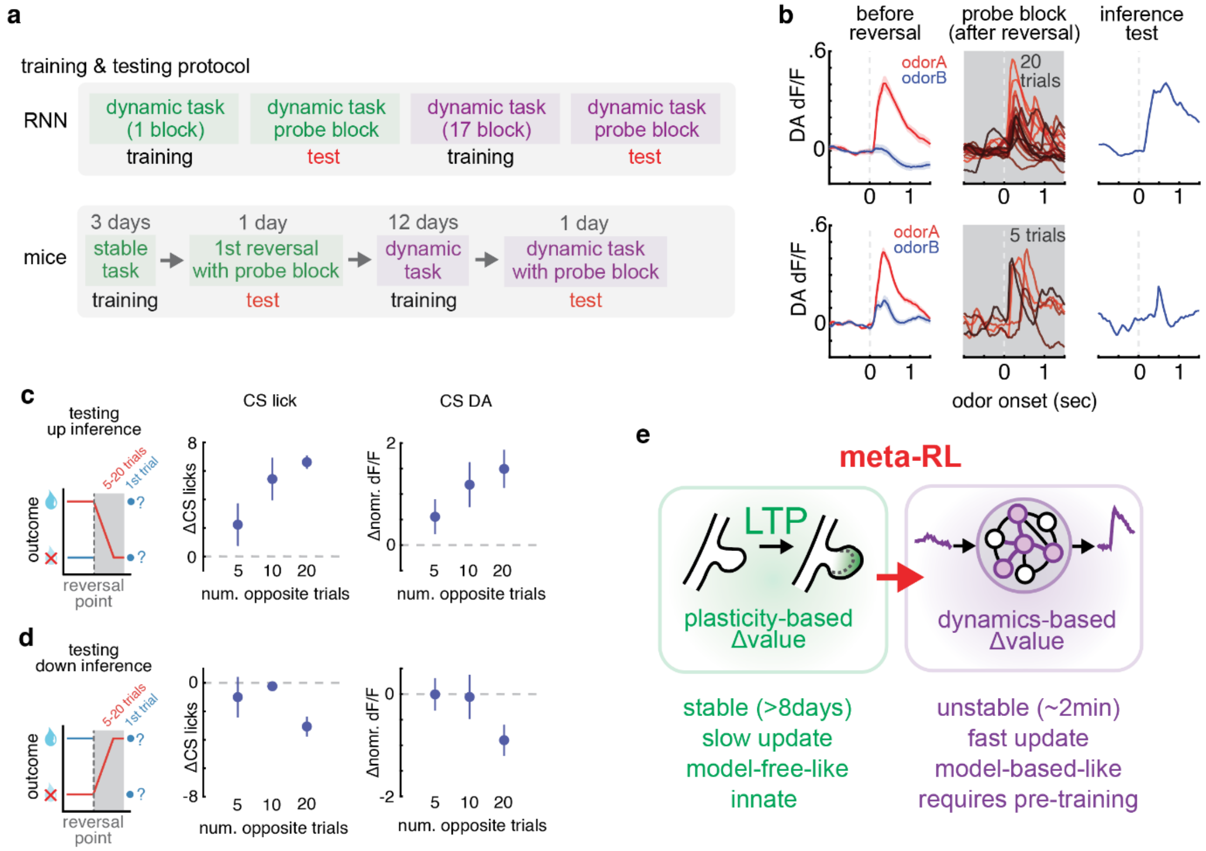
Extended Data Fig. 3 | Histology, detailed behavioral data and alternative models for BLA manipulation experiments. a, Example histology showing a coronal slice and the tip of the injection site. Slice shows DAPI (blue) and DiI (orange) which was mixed with saline or KN-93 to locate the injection site (see Methods). **b,** Summary plot showing all the injection sites for the KN-93/saline infusion experiment (see **Fig. 3a-b**) aligned to the reference atlas. Each dot (orange) represents the injection site for one mouse. **c,** Lick rate for CS+ (red) and CS-(blue) aligned to odor onset after saline infusion (left) or KN-93 infusion (right) (stable/saline, n=11 mice; stable/KN-93, n=8 mice). **d,** Similar plot as in **b** for dynamic task. **e,** Similar plot as in c for the dynamic task (n=7). **f,** Two models consistent with the experimental results of blocking plasticity in the BLA. In model 1, plasticity in the BLA drives value update in the stable task (top, left), but transitions into dynamics-based value update (top, right) after becoming expert in the dynamic task. In model 2, plasticity in BLA drives value update in the stable task (bottom, left), but the locus of plasticity transitions from BLA to another region (region X). Both models can explain why blocking synaptic plasticity in the BLA in the stable task impairs performance but does not impair performance in the dynamic task. A prediction of model 2 is that BLA activity should not be necessary to perform the dynamic task. **g,** Example histology of an emx1-Cre × gtACR1 mouse. Slice shows DAPI (blue) and gtACR1/mCherry (red).



Extended Data Fig. 4 | Distinct timescale of forgetting of stable vs dynamic value in the hybrid task. a, Schematic showing two sessions separate by a 1-day break (grey box). Blue/red lines indicate the outcome for stable cues (CS+/CS-) and purple line indicates the outcome for dynamic cue (CS+/-). Session N+1 starts with the same reward contingency as the end of session N. **b,** Quantification of CS lick rate for stable and dynamic cues for the last 10 trials of session N, and first 10 trials of session n+1. CS lick rate for stable cues remained similar to previous day's rate whereas CS lick rate to the dynamic cue increased relative to its previous level.



Extended Data Fig. 5 | Further quantifications of context coding in the BLA. a, Two models for dynamic value computation in the BLA. In model 1 (left), dynamic value is computed outside BLA (shown here as unknown region X). BLA inherits dynamic value from region X. Context coding units are not present in BLA. In model 2 (right), context coding is either computed locally or inherited from region X. BLA locally computes dynamic value using context information (purple=dynamic value coding neurons; pink=context coding neurons). **b,** Firing rate of all amygdala positive context units (positive CTX, n=59, left) or negative context units (negative CTX, n=171, right) aligned to odor onset. Firing rate is shown for odor C in reward blocks (purple) or non-rewarded blocks (grey). **c,** Within all CTX units in the amygdala, fraction of units that are SV coding (red), DV coding (blue), or SVDV (grey). **d,** Distinct predictions for context update (Δcontext) when coding is reward coding or reward rate coding. Predictions are shown for positive context coding neurons (pos. context, blue) and negative context coding neurons (neg. context, yellow), in either reward blocks (purple) or non-reward blocks (green).



EDF 6 | Further quantification of value inference in RNNs and mice. a, Training and test protocol for value inference related to Main Fig. 6. RNNs were first trained on the 1st block of the dynamic task (equivalent to stable task). On the first reversal point, RNNs were tested for the ability to infer value (probe block). We define this point as naïve RNNs (see main Fig. 6). RNNs were further trained for 17 blocks in the dynamic task, and then tested again on for value inference using the probe block. We define this point as expert RNNs (see main Fig. 6). **b,** Two example sessions showing dopamine photometry signal from a mouse being tested for inference when the opposite trial was presented 20 trials (top) or 5 trials (bottom). *left*, mean traces for odor A and odor B before the reversal (last 10 trials for each odor). *center*, dopamine traces from trials where only one odor was presented consecutively. Color indicates the trial order (light red->dark red=early->late trials). *right*, inference trial where the target cue is presented for the first time (see **Main Fig. 6f, g**). **c,** Up inference. *left,* Schematic showing the protocol for testing upward inference. After reversal, 5-20 trials of one cue were presented serially and then the other cue was presented the first time. Upward inference was quantified as the change in value (CS licks or CS DA) relative to baseline level, defined as the mean of the last 5 trials before reversal (see Methods). *middle,* Quantification of change in CS licks (ΔCS licks) as a proxy for inferred value for different number of opposite trials (5, 10 and 20). *right,* similar quantification for dopamine response (Δnormalized dF/F) (n=6 mice). **d,** Similar quantification as in **c** but for downward inference (n=6 mice). **e,** Summary diagram showing the transition from plasticity to dynamics-based value update for meta-reinforcement learning.