SUSTAINED DOPAMINERGIC PLATEAUS AND NORADRENERGIC DEPRESSIONS BIAS TRANSITIONS INTO EXPLOITATIVE BEHAVIORAL STATES

Aaron C. Koralek[1,2] & Rui M. Costa[1,2]

[1]Zuckerman Mind Brain Behavior Institute

Columbia University, New York, NY

[2] Champalimaud Neuroscience Programme

Champalimaud Centre for the Unknown, Lisbon, Portugal

ABSTRACT

We are constantly faced with the trade-off between exploiting past actions with known outcomes and exploring novel actions whose outcomes may be better. The balance between exploitation and exploration has been hypothesized to rely on multiple neuromodulator systems, namely dopaminergic neurons of the substantia nigra pars compacta (SNc) and noradrenergic neurons of the locus coeruleus (LC). However, little is known about the dynamics of these neuromodulator systems during exploitative and exploratory states, or how they interact. We developed a novel behavioral paradigm to capture exploitative and exploratory behavioral states, and imaged calcium dynamics in genetically-identified dopaminergic SNc neurons and noradrenergic LC neurons during the transitions between these states. We found dichotomous changes in sustained activity in SNc and LC during exploitative bouts of action-reward, with SNc showing higher and LC showing lower sustained activity. Exploitative states were also marked by a lengthening of positive SNc response plateaus and negative LC response depressions, as well as hysteretic dynamics in SNc networks. Chemogenetic enhancement of dopaminergic and noradrenergic excitability favored exploitative and exploratory states, respectively. Together, these data suggest that opponent changes in dopaminergic and noradrenergic activity states modulate the transitions between exploitative and exploratory behavioral states, with important implications for downstream circuit dynamics.

INTRODUCTION

At any given moment, animals must choose their next action from a vast repertoire of possible behavioral responses. Some actions have been performed repeatedly in the past and

therefore have well-known outcomes, while others have less certain but potentially better outcomes. In addition, there are fluctuations in the motivational drive to perform some actions over others, depending on the current state of both the environment and the animal. This trade-off between exploiting past actions and exploring novel ones gives humans and animals the amazing ability to explore new environments and develop novel behavioral responses following environmental changes. This balance has been proposed to rely on midbrain dopaminergic neurons of the substantia nigra pars compacta (SNc)[1] and noradrenergic neurons of the locus coeruleus (LC)[2-3].

A role for dopamine (DA) in low-level motor variability has been reported[4-6], and a growing body of work suggests that DA also affects variability in action selection[7]. SNc neurons fire strongly during the initiation of self-paced action sequences[8] and patients with PD exhibit deficits in action initiation[9], suggesting a role for DA in the volitional initiation of learned actions. In addition, deficits in choice reversal learning[10] and attentional set switching[11] have been demonstrated following DA depletion. Computational modeling work has also predicted a central role for DA signaling in modifying the action selection probabilities associated with two-choice tasks[12]. Similarly, recent work suggests that levels of norepinephrine (NE), and specifically activity of the noradrenergic projections to prefrontal cortices, modulate levels of stochastic responding in rodents[13-14]. Both dopaminergic and noradrenergic systems therefore appear to play a role in the balance between exploitative and exploratory responding.

Although past work has studied isolated exploitative and exploratory choices[22-24] and computational modeling work has predicted a role for catecholamines in these decisions[12], the majority of this work has focused on single-trial decisions, thus obscuring the longer-term state changes that define exploitative and exploratory states of action selection. These exploitative

states, characterized by engaged and motivated performance of a well-learned skill to achieve a desired outcome, might be similar to what is colloquially referred to as "being in the zone" or the "hot hand effect"[25-26].

We developed a novel behavioral paradigm in mice that probes action selection among many possible actions over long time scales and allows us to bias behavior towards exploitative or exploratory states using environmental change. This paradigm permitted us to the study the behavior of animals away from ceiling or floor performance, and hence to study the emergence of bouts of exploitative or exploratory choices. We imaged the activity of populations of individual dopamine neurons of the sustantia nigra pars compacta (SNc) and noradrenergic neurons of the locus coeruleus (LC) and found striking changes in sustained dopaminergic and noradrenergic activity that cumulatively emerge when animals are in exploitative behavioral states, repeatedly performing well-known skills to achieve desired outcomes. These exploitative states are marked by lengthened response plateaus and hysteretic network dynamics in SNc neurons, as well as lengthened response depressions in LC neurons. Finally, we induced sustained changes in the excitability of SNc dopaminergic neurons and LC noradrenergic neurons and found that this biased the balance between exploitative and exploratory behavioral states.

RESULTS

A Novel Task for Probing Exploitative and Exploratory Behavioral States

Theoretical work suggests that novel or unstable environments enhance exploration, while stable or known environments favor exploitation[27]. This suggests that animals' choices will be more exploratory following an unexpected change in the environmental reward structure and choices will become exploitative as the animals become acquainted with the new structure. If the

probabilities of selecting a range of actions are known, then the entropy of this distribution captures levels of exploitative and exploratory behavior.

To develop a framework for studying exploitative and exploratory states in mice, we created a nose poke sequence task in which mice can choose between many possible actions. Mice were placed in an operant chamber with 3 equidistant nose pokes (Fig. 1a). A sequence of 3 pokes in a specific order was rewarded. Importantly, mice were given no trials and few cues to guide learning, but instead had to actively explore the environment to determine the reward structure. When mice performed the target sequence, reward was supplied via a central reward port. There are 27 possible sequences, providing a broad distribution of potentially selectable actions.

Performance improved significantly over the course of roughly one month of training, as seen both in an increase in reward rate (Fig. 1b) and an increase in the proportion of pokes that compose the rewarded sequence relative to total pokes (Fig. 1c). Chance levels of performance were assessed by modeling an agent that performs the same number of pokes as the mice on each day, but selects each poke randomly (Fig. 1b-e, red lines). Animals performed significantly above this assessment of chance levels for all behavioral measures. Importantly, after training, mice were well above floor performance but also below ceiling performance, allowing us to study the transitions between exploitative and exploratory bouts. Over the course of training, we observed a decrease in the entropy of the animals' selected sequences (Fig. 1d), as well as a decrease in the entropy of the animals' transitions between nose pokes (Fig. 1e), suggesting that animals are initially sampling a relatively wide range of possible actions, but gradually refine these choices to focus more on the rewarded sequence. When examining the number of pokes at response ports between checks for reward at the reward port ("Inter-Check Interval"), we observed that animals check for reward after a majority of response pokes in early learning, but begin to chunk behavior

into groups of three nose pokes in late learning (Fig. 1f), suggesting that they have learned to mostly perform three-poke sequences and understand the structure of the task. The time that mice took to perform the rewarded sequence also decreased significantly with training (Fig. 1g).

When the animals were proficient at performing a target sequence, we changed the target sequence to be rewarded. This sequence change occurred within a behavioral session, and we found that the prevalence of the previously-rewarded sequence decreased following the sequence change (Fig. 1h, top, blue) and the entropy of selected sequences increased (Fig. 1h, bottom). The prevalence of the newly-rewarded sequence gradually increased as mice discovered the new reward structure (Fig. 1h, top, red) and the entropy returned to lower levels (Fig. 1h, bottom). Across all animals, we observed a significant decrease in performance (Fig. 1i) and an increase in entropy (Fig. 1j) immediately following the rule change, suggesting that we were successfully driving animals into a more exploratory behavioral state by changing the reward structure of the environment.

Sustained Dopaminergic and Noradrenergic Modulations During Exploitative States

We next imaged calcium dynamics in genetically-identified dopaminergic and noradrenergic cells of the SNc and LC, respectively, through chronically-implanted gradient index (GRIN) lenses (Fig. 2a). We first examined phasic bursting in the populations before and after a change in the rewarded sequence (Fig. 2b), with a focus on three conditions. Namely, “Exploit” designates the epoch before the rule change when mice were exploiting a well-known reward structure, with PSTHs time-locked to performance of the target sequence. The “Explore” conditions, in contrast, designate the epoch after the rule change when mice were exploring a novel reward structure, and this is subdivided into two conditions. “Explore - Old” represents trials when

mice exhibited perseverative errors during the exploratory epoch and performed the previously-rewarded action that is no longer rewarded, and Explore-Old therefore includes the same action as Exploit, but a different outcome. Conversely, “Explore - New” represents trials when mice performed the newly-rewarded action, and Explore-New therefore includes the same outcome as Exploit, but a different action. In both regions, we observed qualitatively similar phasic responses to rewards during exploitative and exploratory epochs. However, these phasic responses appeared to begin from different baseline levels of activity during exploitation and exploration, with SNc baseline activity enhanced during exploitation and LC baseline activity reduced during exploitation (Fig. 2b). In SNc, we noted that exploitative and exploratory rewards result in comparable peak magnitudes, despite the change in baseline sustained activity, which is consistent with recent reports suggesting that reward expectation is marked by increases in baseline activity rather than decreases in peak amplitude[28]. When we expanded the time axis, we found that these baseline changes develop leading up to reward and fade afterwards, lasting for a total of approximately 60 seconds surrounding exploitative rewards (Fig. 2c).

We first asked whether these baseline changes were due to the averaging of many brief, staggered bursts of activity or were instead a slowly-varying component of the activity. To disambiguate between these possibilities, we preprocessed our calcium data in a manner that separates quickly-varying and slowly-varying components of the signal (Fig. 2d and Methods) and we then used these component traces to create peri-stimulus time histograms (PSTHs) that are analogous to those we created with the full calcium traces. For both SNc and LC, we found that the changes in baseline activity were not observed when using only the quickly-varying components (Fig. 2d, right, red), suggesting that the observed baseline shifts are not due to the

averaging of many brief, jittered bursts. However, the sustained effects were still present when using only the slowly-varying components of the activity (Fig. 2d, right, blue).

We next asked whether these sustained activity changes were due simply to differences in reward rate during exploitation and exploration. We therefore ran animals on a version of the task in which all possible three-poke sequences are rewarded with either high probability ("Day H"; 80%) or low probability ("Day L"; 20%). We did not observe changes in baseline activity in either SNc or LC on either Day H or Day L (Fig. 2e). Importantly, the reward rate on Day H was comparable to that during exploitation, but the entropy was significantly higher (Supp. Fig. 1), suggesting that the baseline shifts are more closely related to changes in choice entropy than they are to changes in reward rate. In addition, we did not observe baseline activity changes in early learning (Supp. Fig. 2) or in the ventral tegmental area dopaminergic neurons (VTA; Supp. Fig. 3). This phenomenon was also not the result of averaging across neurons. We found that roughly 20-25% of individual neurons in SNc and LC exhibit these sustained changes in baseline activity, with clearly distinct activity profiles relative to other neurons in the network during exploitative states or to all neurons during exploratory states (Fig. 2f).

Due to the slowly-varying nature of these baseline changes, we wondered whether they could be representing behavioral variables in a reinforcement learning (RL) context. We therefore fit a basic RL model to our behavioral data to extract estimates of action value (Q), state value (V), and reward prediction error (RPE). We found that the overall peri-event correlation of activity in all neurons and the different RL variables was strikingly low (Supp. Fig. 4a-b). In addition, we constructed reward-locked PSTHs using the estimates of Q, V, and RPE, and found these to exhibit very little activity at time points distant from reward (Supp. Fig. 4c). Finally, we correlated the full time-courses of Q, V, and RPE with smoothed fluorescence traces from SNc and found these

correlations to also be very weak (Supp. Fig. 4d). We found, as expected, that individual cells vary greatly in the strength and direction of their correlations with these RL parameter estimates, with some cells showing strong positive correlations and others showing negative correlations (Fig. 2g). We therefore asked whether cells that are positively correlated with these RL parameters exhibited different levels of sustained activity than cells that are negatively correlated with these parameters, and found there to be very little difference in the sustained reward-locked activity of neurons correlated with estimates of either action value, state value, or RPE (Fig. 2h). Finally, we found no strong relationship between the cells that we classified as exhibiting sustained effects (in Fig. 2f) and the distribution of their correlations with these RL parameter estimates (mean±SD, Q: 0.198±0.295, V: 0.142±0.299, RPE: 0.059±0.249). Together, these results suggest that, although the activity of many cells in the network was correlated with classic RL parameters, the observed changes in sustained activity cannot be accounted for simply by slowly-changing estimates of action and state value.

Sustained Activity Cumulatively Emerges in Dopaminergic and Noradrenergic Networks During Reward Bouts

Although we found that the sustained activity changes were not due to average reward rate (Fig. 2e), we investigated whether the structure of "target action-reward" events changed following the sequence change. We therefore performed a "reward autocorrelation", where time 0 indicates the occurrence of a reward and the occurrence of other rewards is averaged time-locked to this point. We found no change in the overall temporal structure of reward occurrence before or after the sequence change (Fig. 3a).

We therefore asked whether there was instead a change in the neural responses to rewards. To address this, we defined "target action-reward bouts" (hereafter "action-reward bouts") as clusters of action-reward pairs that are separated from each other by less than 10 seconds and separated from other action-reward pairs by more than 20 seconds, and we then averaged activity based on the action-reward pair's position within a bout (Fig. 3b). Because the inter-event timing within action-reward bouts is not required to be constant, we investigated whether the number of action-reward bouts changed across behavioral states, despite previously observing no shift in the overall temporal pattern of action-reward events (Fig. 3a). We found that the rate of occurrence of action-reward bouts increased significantly in exploitative relative to exploratory states (Fig. 3c), suggesting that action-reward bouts could be an important characteristic of the transitions between these behavioral states. We therefore investigated neural responses throughout these bouts during exploitation and we found that dopaminergic activity accumulated over the course of a bout and returned to low levels by the end of the bout (Fig. 3c, top, and Supp. Fig. 5), while LC activity decreased consistently over the course of the bout (Fig. 3c, bottom). These patterns were not present during exploration (Fig. 3d), on Day H (Fig. 3e), or on Day L (Fig. 3f). Importantly, reward bouts are common on Day H, but they are preceded by different actions. Intriguingly, the base-to-peak of activity within an action-reward bout remains fairly constant throughout the bout (Fig. 3c, upper inset). However, this measure decreases significantly over the course of a bout, similar to the classic effects of RPE on DA cell activity[29], when performing the analysis using traces that contain only quickly-varying components of the signal to mimic data preprocessing methods in which baseline activity is corrected on a short timescale (Fig. 3c, lower inset). This is in-line with recent proposals that, with increased reward predictability, DA cells exhibit enhanced baseline activity rather than decreased bursting activity[28].

We next asked whether the activity profile we observed during action-reward bouts was sufficiently characteristic of these bouts to enable prediction of bout occurrences from neural activity. We found that prediction of action-reward bouts by single neurons was significantly better than chance (Supp. Fig. 6). Furthermore, if we considered only neurons whose activity was most predictive of action-reward bouts, we found that these predictive neurons exhibited stronger sustained changes in activity surrounding exploitative action-reward pairs than the rest of the population in both SNc and LC (Fig. 3g). Together, these data suggest that sustained activity accumulates positively in SNc and negatively in LC as animals perform bouts of the target sequence in exploitative, but not exploratory, states. Furthermore, these bouts define an exploitative behavioral state, whereby animals are engaged in repeatedly performing a well-known skill to achieve a favorable outcome.

Altered Neuronal Response Dynamics Drive Sustained Activity

We next asked what neuronal response differences during exploitative and exploratory states could produce the distinct ways in which activity accumulates over the course of action-reward bouts to produce sustained activity shifts. We therefore quantified the average length of all positive and negative response transients during exploitative and exploratory epochs. We found an increase in the duration of positive response transients in SNc neurons during exploitative relative to exploratory behavioral states, resulting in response plateaus (Fig. 4a). There was no concomitant change in the duration of negative response transients. The duration of positive response transients in SNc neurons during exploratory states was similar to that observed during early learning (Supp. Fig. 2), on Day H or Day L (Supp. Fig. 1), or in the VTA (Supp. Fig. 3). In contrast, in LC neurons, we found no change in the duration of positive response transients across these behavioral states,

but the duration of negative response transients was significantly longer in exploitative relative to exploratory states, resulting in response depressions (Fig. 4c). To examine whether these activity changes in individual neurons were also reflected in changing network interactions, we created cross-correlation histograms, where activity from all cells was time-locked to large fluorescence bursts in other simultaneously-recorded cells. Importantly, in SNc during exploitative states, we found that cells in the network tend to increase activity together, but then continue to fire afterwards, exhibiting network-level hysteretic effects (Fig. 4b). In LC, correlated activity was generally increased during exploitation, but the shape of this response was unchanged (Fig. 4l). The asymmetry seen in SNc during exploitative states is also not present in early learning (Supp. Fig. 2), on Day H or Day L (Supp. Fig. 1), or in the VTA (Supp. Fig. 3). Dopaminergic and noradrenergic networks therefore exhibit striking changes in response dynamics across exploitative and exploratory behavioral states.

To investigate whether these changing response dynamics could produce the observed changes in sustained activity, we created reward convolution traces by convolving an impulse response function (IRF) with the occurrences of rewards in our behavioral data (Fig. 4e-f). For SNc, this IRF was a simple exponential of varying length (Fig. 4e), while for LC, this IRF was created by smoothing the average population response to unexpected rewards (Fig. 4f). As we increased the duration of the IRFs, we observed the emergence of sustained activity surrounding reward that matches that observed during exploitative states in SNc (Fig. 4h,j). This was not observed if we more closely match the IRF exponential duration to the statistics of our neural data in baseline settings (Fig. 4g,i). However, for IRFs matched to our neural data, the addition of hysteretic network dynamics to the model results in model responses nearly identical to those seen in our data during exploitative states (Fig. 4h, inset). The LC IRF, on the other hand, is biphasic,

with both a positive and negative phase. Intriguingly, in baseline settings, this IRF is asymmetric, with the negative phase of the response lasting 1.5 times the duration of the positive phase (Fig. 4f), suggesting the possibility of an innate bias towards negative activity accumulations in the LC network. If we then alter the length of the negative phase of this IRF while holding the positive phase constant in our reward convolution model, we again see the emergence of sustained changes in activity with longer IRFs (Fig. 4j). This suggests that the observed modifications to the dynamics of plateau and depression responses in either SNc or LC can recapitulate the observed changes in baseline activity that we see during exploitative behavioral states. Importantly, the amplitude of positive and negative transients does not change between exploitative and exploratory states (Supp. Fig. 7), demonstrating that the sustained activity changes cannot be accounted for by differences in the amplitude of transients. For both SNc and LC, if we look across all conditions (Exploit, Explore - Old, and Explore - New), we see that these extended IRFs produce analogous changes in baseline activity in response to both exploitative (Exploit) and exploratory (Explore - New) rewards (Fig. 4k-l), which is inconsistent with our data (Fig. 2c). Flexible transitions between these distinct regimes of neuronal response dynamics are therefore necessary to produce the neuronal response patterns that we observed across exploitative and exploratory behavioral states.

Increasing Dopaminergic or Noradrenergic Excitability Biases Transitions between Exploitative and Exploratory States

We next asked whether these sustained changes in baseline activity levels play a causal role in modulating the transitions between exploitative and exploratory states. Because these baseline shifts were due to changes in neural response dynamics that accumulate over the course of exploitative bouts, we used chemogenetic manipulations to modulate the excitability of

dopaminergic and noradrenergic populations, rather than using optogenetic manipulations that drive excitability briefly. We used the designer receptor exclusively activated by designer drugs (DREADD) hM3Dq to enhance excitability in genetically-identified dopaminergic or noradrenergic populations[30-31].

Animals expressing either hM3Dq or mCherry were trained on the task and given intraperitoneal injections of either clozapine-N-oxide (CNO; the hM3Dq ligand) or vehicle (VEH) during exploitative states, exploratory states, and on Day H (Fig. 5a). During exploitative states, we found that enhancing LC excitability with CNO produced an increase in transition entropy and a decrease in reward rate relative to VEH, and this effect was not seen in control animals expressing mCherry in LC (Fig. 5c,e). The same LC manipulation during exploratory states, when animals still do not know the correct action sequence, did not produce any effect, showing that this was not a general effect of LC manipulations on behavior. We found no change in transition entropy or reward rate when enhancing SNc excitability during exploitative states, potentially because animals were already in exploitative states and could not be pushed further in performance (Fig. 5b and Supp. Fig. 8). Similarly, we found no difference when enhancing SNc excitability during exploratory states, before animals had learned which action sequence would lead to reward. We therefore tested the same manipulation on Day H, when animals could perform a wider range of actions to get the same rate of reward. We found that enhancing SNc activity with CNO produced marked changes in the structure of their responses. A significantly higher proportion of the total rewards were in action-reward bouts following CNO injection relative to VEH, and this effect was not present in animals expressing mCherry in SNc (Fig. 5b,d). The mean number of action-reward pairs per action-reward bout was also increased significantly following CNO (Fig. 5b,d). This restructuring of action-reward bouts following CNO injections in animals expressing hM3Dq in

SNc was not seen in animals expressing hM3Dq in LC (Supp. Fig. 8). Enhanced noradrenergic activity therefore appears to increase levels of response entropy and disrupt performance of a well-learned skill, driving transitions into exploratory behavioral states, while enhanced dopaminergic activity results in the restructuring of task performance into exploitative action-reward bouts.

DISCUSSION

We developed a novel behavioral framework in rodents to capture exploitative and exploratory states of action selection, and we found marked changes in sustained activity in both dopaminergic and noradrenergic populations across these states. We found these sustained baseline changes to be due to enhanced response plateaus and depressions during exploitative states in both SNc and LC, as well as hysteretic network dynamics in SNc, that resulted in accumulations of activity during exploitative bouts, when animals repeatedly perform well-learned skills to achieve desired outcomes. Increasing neuronal excitability in dopaminergic and noradrenergic populations changed the structure of behavior in accordance with these theorized behavioral states. Together, our results suggest that sustained plateaus in dopaminergic networks and sustained depressions in noradrenergic networks increase the likelihood of transitioning into an exploitative state: a state of inspired engagement in performing a well-known skill that might be colloquially referred to as “being in the zone”.

Intermediate-Scale Behavioral States

The design of our task and experiments allowed us to investigate behavioral and neural states that occur on an intermediate-timescale between the scale of synaptic signaling (milliseconds) and the scale of long-term potentiation (hours to days), a timescale that is much

more similar to that experienced during ongoing behavior and perhaps more relevant to slower-acting neuromodulator systems. By expanding the time axis out to multiple minutes, we were able to identify these intermediate-scale neural effects, as well as the enhanced responses to exploitative bouts that exist on these intermediate timescales. Further work is necessary to more fully characterize the composite changes to widespread neural circuits that accompany and define these intermediate-scale behavioral states.

Differential Downstream Effects of Sustained Versus Brief Changes

There are a number of ways that the sustained changes we found in dopaminergic and noradrenergic activity could impact downstream circuits for action selection in the dorsal striatum (DS), in the case of DA, and the anterior cingulate cortex (ACC), in the case of NE. Both systems are known to contain a range of receptor subtypes with distinct postsynaptic effects[32-35]. In the DS, the main DA receptor types, D1R and D2R, are expressed in segregated neuronal populations[32-33], and modeling work has suggested that D1Rs respond preferentially to brief fluctuations in DA, while D2Rs track more sustained changes[34]. Accordingly, recent work has demonstrated distinct effects on prefrontal dynamics when dopaminergic networks are stimulated on different temporal scales[36]. Similarly, ACC cells contain both α1 and α2 NE receptors[35]. While α2Rs are high affinity and respond at low rates of NE release, α1Rs are low affinity and are only engaged at high rates of release[37]. Both systems therefore contain receptors tuned to either brief or sustained changes in neuromodulator levels. These receptors are also associated with distinct behavioral effects[32-37], suggesting that brief and sustained changes in DA and NE could have profoundly different effects on downstream network dynamics and, in turn, the behavioral state of the animal. These possibilities are currently being investigated.

Potential Biological Substrates of Response Plateaus and Depressions

The response plateaus and depressions observed in the current work using calcium imaging have a number of possible biological underpinnings. Most simply, these enhanced response functions and network dynamics could reflect a change in the excitability or gain of individual neurons, which would be consistent with the observed behavioral effects following chemogenetic manipulations. This could, as one example, involve a change in the probability of neuronal up- or down-states across the population[44]. Alternatively, these plateaus and depressions could reflect changing lateral interactions within the network, which is also consistent with the observed hysteretic network effects[45]. More granular investigations with physiological methods are necessary to disentangle these possibilities.

Relationship to Pathological States

Our results could have important implications for a range of mental disorders and maladaptive behavioral states marked by aberrant action selection, such as obsessive-compulsive disorder[15-16], schizophrenia[17-18], and addiction[19-21], among others[38-40]. These syndromes can all be viewed as disorders of exploitative or exploratory responding, and our results suggest that alterations to the dopaminergic and noradrenergic systems could be part of the pathological mechanism or useful for managing symptomatology. In addition, our results suggest the possibility that modulations of the dopaminergic and noradrenergic systems could alleviate perseverative states in non-pathological populations, or even enhance creativity in the search for novel behaviors. These data therefore point to a number of therapeutic targets worthy of further investigation.

Conclusion

In summary, we developed and validated a novel behavioral paradigm for probing exploitative and exploratory behavioral states, and we found that neuronal populations in SNc and LC exhibited marked changes in sustained activity levels across these two states. These sustained effects were due to enhanced response plateaus and hysteretic population activity in SNc, as well as enhanced response depressions in LC, that both produced accumulations of activity during exploitative bouts of responding. Artificially enhancing dopaminergic or noradrenergic excitability modulated the likelihood of transitioning into these exploitative states, providing causal evidence for the role of sustained response properties in shaping behavior. Together, these results clarify the neural processes that modulate our choice of behaviors to either maintain a series of successes and capitalize on our learned skills or, conversely, to explore alternative actions and discover novel, creative behavioral responses to a complex and nuanced environment.

REFERENCES

1. Costa, R.M. (2007). Plastic corticostriatal circuits for action learning: what's dopamine got to do with it? *Ann NY Acad Sci*, 1104:172-91.
2. Aston-Jones, G. & Cohen, J.D. (2005). An integrative theory of locus coeruleus-norepinephrine function: adaptive gain and optimal performance. *Annu Rev Neurosci*, 28:403-450.
3. Usher, M., Cohen, J.D., Servan-Schreiber, D., Rajkowski, J., & Aston-Jones, G. (1999). The role of locus coeruleus in the regulation of cognitive performance. *Science*, 283:549-554.
4. Budzillo, A., Duffy, A., Miller, K.E., Fairhall, A.L., & Perkel, D.J. Dopaminergic modulation of basal ganglia output through coupled excitation-inhibition. *Proc Natl Acad Sci*, 114(22):5713-5718.
5. Murugan, M., Harward, S., Scharff, C., & Mooney, R. (2013). Diminished foxP2 levels affect dopaminergic modulation of corticostriatal signaling important to song variability. *Neuron*, 80(6):1464-76.
6. Galea, J.M., Ruge, D., Buijink, A., Bestmann, S., & Rothwell, J.C. (2013). Punishment-induced behavioral and neurophysiological variability reveals dopamine-dependent selection of kinematic movement parameters. *J Neurosci*, 33(9):3981-8.
7. Costa, R.M. (2011). A selectionist account of de novo action learning. *Curr Opin Neurobiol*, 21(4):579-86.
8. Jin, X. & Costa, R.M. (2010). Start/stop signals emerge in nigrostriatal circuits during sequence learning. *Nature*, 466(7305):457-62.
9. Benecke, R., Rothwell, J.C., Dick, J.P., Day, B.L., & Marsden, C.D. (1987). Disturbance of sequential movements in patients with Parkinson's disease. *Brain*, 110:361-79.
10. Izquierdo, A., Wiedholz, L.M., Millstein, R.A., Yang, R.J., Bussey, T.J., Saksida, L.M., & Holmes, A (2006). Genetic and dopaminergic modulation of reversal learning in a touchscreen-based operant procedure for mice. *Brain Behav Res*, 171:181-88.
11. Klanker, M., Feenstra, M., & Denys, D. (2013). Dopaminergic control of cognitive flexibility in humans and animals. *Front Neurosci*, 7:201.
12. Humphries, M.D., Khamassi, M., & Gurney, K. (2012). Dopaminergic control of the exploration-exploitation trade-off via the basal ganglia. *Front Neurosci*, doi: 10.3389/fnins.2012.00009.
13. Tervo, D.G.R., Proskurin, M., Manakov, M., Kabra, M., Vollmer, A., Branson, K., & Karpova, A.Y. (2014). Behavioral variability through stochastic choice and its gating by anterior cingulate cortex. *Cell*, 159(1):21-32.
14. Karlsson, M.P., Tervo, D.G., & Karpova, A.Y. (2012). Network resets in medial prefrontal cortex mark the onset of behavioral uncertainty. *Science*, 338(6103):135-9.
15. Pauls, D.L., Abramovitch, A., Rauch, S.L., & Geller, D.A. (2014). Obsessive-compulsive disorder: an integrative genetic and neurobiological perspective. *Nat Rev Neurosci*, 15:410-424.
16. Koo, M.S., Kim, E.J., Roh, D., & Kim, C.H. (2010). Role of dopamine in the pathophysiology and treatment of obsessive-compulsive disorder. *Expert Rev Neurother*, 10:275-290.
17. Strauss, G.P., Frank, M.J., Waltz, J.A., Kasanova, Z., Herbener, E.S., & Gold, J.M. (2011). Deficits in positive reinforcement learning and uncertainty-driven exploration are associated with distinct aspects of negative symptoms in schizophrenia. *Biol Psychiatry*, 69:424-431.

18. Grace, A.A. (2016). Dysregulation of the dopamine system in the pathophysiology of schizophrenia and depression. *Nat Rev Neurosci*, 17:524-532.
19. Keiflin, R. & Janak, P.H. (2015). Dopamine prediction errors in reward learning and addiction: from theory to neural circuitry. *Neuron*, 88(2):247-63.
20. Schultz, W. (2011). Potential vulnerabilities of neuronal reward, risk, and decision mechanisms to addictive drugs. *Neuron*, 69(4):603-17.
21. Berke, J.D. & Hyman, S.E. (2000). Addiction, dopamine, and the molecular mechanisms of memory. *Neuron*, 25(3):515-32.
22. Cohen, J.D., McClure, S.M., & Yu, A.J. (2007). Should I stay of should I go? How the human brain manages the trade-off between exploitation and exploration. *Philos Trans R Soc Lond B Biol Soc*, 362:933-942.
23. Dayan, P. & Daw, N.D. (2008). Decision theory, reinforcement learning, and the brain. *Cogn Affect Behav Neurosci*, 8:429-53.
24. Daw, N.D., Niv, Y., & Dayan, P. (2005). Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioral control. *Nat Neurosci*, 8:1704-11.
25. Neiman, T. & Loewenstein, Y. (2011). Reinforcement learning in professional basketball players. *Nat Commun*, 2:569.
26. Blanchard, T.C., Wilke, A., & Hayden, B.Y. (2014). Hot-hand bias in rhesus monkeys. *J Exp Psychol Anim Learn Cogn*, 40(3):280-6.
27. Speekenbrink, M. & Konstantinidis, E. (2015). Uncertainty and exploration in a restless bandit problem. *Top Cogn Sci*, 7:351-67.
28. Hamid, A.A., Pettibone, J.R., Mabrouk, O.S., Hetrick, V.L., Schmidt, R., Vander Weele, C.M., Kennedy, R.T., Aragona, B.J., & Berke, J.D. (2016). Mesolimbic dopamine signals the value of work. *Nat Neurosci*, 19(1):117-126.
29. Schultz, W., Dayan, P., & Montague, P.R. (1997). A neural substrate of prediction and reward. *Science*, 275(5306):1593-9.
30. Armbruster, B.N., Li, X., Pausch, M.H., Herlitze, S., & Roth, B.L. (2007). Evolving the lock to fit the key to create a family of G protein-coupled receptors potently activated by an inert ligand. *Proc Natl Acad Sci*, 104(12):5163-8.
31. Urban, D.J. & Roth, B.L. (2015). DREADDs (designer receptors exclusively activated by designer drugs): chemogenetic tools with therapeutic utility. *Anny Rev Pharmacol Toxicol*, 55:339-417.
32. Kravitz, A.V., Tye, L.D., & Kreitzer, A.C. (2012). Distinct roles for direct and indirect pathway striatal neurons in reinforcement. *Nat Neurosci*, 15(6):816-8.
33. Tecuapetla, F., Jin, X., Lima, S.Q., & Costa, R.M. (2016). Complementary contributions of striatal projection pathways to action initiation and execution. *Cell*, 166(3):703-715.
34. Dreyer, J.K., Herrik, K.F., Berg, R.W., & Hounsgaard, J.D. (2010). Influence of phasic and tonic dopamine release on receptor activation. *J Neurosci*, 30:14273-83.
35. Berridge, C.W. & Spencer, R.C. (2016). Differential cognitive actions of norepinephrine a2 and a1 receptor signaling in the prefrontal cortex. *Brain Res*, 1641: 189-96.
36. Lohani, S., Martig, A.K., Deisseroth, K., Witten, I.B., Moghaddam, B. (2019). Dopamine modulation of prefrontal cortex activity is manifold and operates at multiple temporal and spatial scales. *Cell Rep*, 27(1):99-114.
37. Arnsten, A.F. (2000). Through the looking glass: differential noradrenergic modulation of prefrontal cortical function. *Neural Plast*, 7(1-2):133-46.

38. Nutt, D.J. (2006). The role of dopamine and norepinephrine in depression and antidepressant treatment. *J Clin Psychiatry*, 67:3-8.
39. Ordway, G.A., Schenk, J., Stockmeier, C.A., May, W., & Klimek, V. (2003). Elevated agonist binding to α2-adrenoceptors in the locus coeruleus in major depression. *Biol Psychiatry*, 53:315-323.
40. Tye, K.M., Mirzabekov, J.J., Warden, M.R., Ferenczi, E.A., Tsai, H.C., Finkelstein, J., Kim, S.Y., Adhikari, A., Thompson, K.R., Andalman, A.S., Gunaydin, L.A., Witten, I.B., & Deisseroth, K. (2013). Dopamine neurons modulate neural encoding and expression of depression-related behaviour. *Nature*, 493(7433):537-41.
41. Desrochers, T.M., Jin, D.Z., Goodman, N.D., & Graybiel, A.M. (2010). Optimal habits can develop spontaneously through sensitivity to local cost. *Proc Natl Acad Sci*, 107(47):20512-7.
42. Zhou, P., Resendez, S.L., Rodriguez, Romaguera, J., Jimenez, J.C., Neufeld, S.Q., Giovanucci, A., Friedrich, J., Pnevmatikakis, E.A., Stuber, G.D., Hen, R., Kheirbek, M.A., Sabatini, B.L., Kass, R.E., & Paninski, L. (2018). Efficient and accurate extraction of in vivo calcium signals from microendoscopic video data. *Elife*, doi: 10.7554/eLife.28728.
43. Pnevmatikakis, E.A., Soudry, D., Gao, Y., Machado, T.A., Merel, J., Pfau, D., Reardon, T., Mu, Y., Lacefield, C., Yang, W., Ahrens, M., Bruno, R., Jessell, T.M., Peterka, D.S., Yuste, R., &Paninski, L. (2016). Simultaneous denoising, deconvolution, and demixing of calcium imaging data. *Neuron*, 89(2):285-99.
44. UP-DOWN cortical dynamics reflect state transitions in a bistable network. *Elife*, 6: doi: 10.7554/eLife.22425.
45. Sahasranamam, A., Vlachos, I., Aertsen, A., & Kumar, A. (2016). Dynamical state of the network determines the efficacy of single neuron properties in shaping the network activity. *Sci Rep*, 6:26029.

ACKNOWLEDGMENTS

This material is based on work supported by the National Institutes of Health grant (1K99MH118412-01) to A.C.K., the NARSAD Young Investigator grant (PG010634) to A.C.K., the Simons Collaboration on the Global Brain grant (348880) to A.C.K., the Bial Foundation grant (413/14) to A.C.K., the European Research Council Consolidator grant (COG 617142) to R.M.C., and the U19 Brain Initiative grant (5U19NS104649) to R.M.C. We thank Dr. Vivek Athalye and Dr. James Murray for insightful discussion.

COMPETING INTEREST STATEMENT

The authors declare no competing financial interests.

FIGURE LEGENDS

Figure 1. A novel task for probing action selection. a. Task schematic. Mice are presented with three equidistant nose poke ports and must discover a rewarded sequence of three nose pokes in order. There are no trials, but instead a moving buffer of the last 3 nose pokes is monitored for the rewarded sequence. Over the course of training, **b.** reward rate increases, **c.** the prevalence of the rewarded sequence increases, **d.** the entropy of the distribution of selected actions decreases, and **e.** the entropy of the transitions between pokes decreases. Red lines indicate estimates of chance performance. **f.** Early in learning, mice check for reward after every response poke, but late in learning they begin to make 3 response pokes before checking for reward, suggesting that they are adapting their behavior to the task statistics. **g.** The average time necessary to complete the rewarded sequence decreases with training. **h**. Animals begin reversal sessions exploiting a well-known sequence and the rewarded sequence is changed during the session. Top: Prevalence of the sequence that was rewarded before (blue) or after (red) the change in rewarded sequence (dotted line). Bottom: The entropy of selected actions rises after the change in rewarded sequence (dotted line) and falls again as animals discover the new rewarded sequence. **i.** Reward rate decreases and **j.** entropy increases following a change in the rewarded sequence. Error bars denote s.e.m.

Figure 2. Sustained dopaminergic and noradrenergic activity modulations. a. Schematic of endoscope imaging. **b-c.** PSTHs time-locked to exploitative rewards before the sequence change ("Exploit"; black), perseverative errors after the sequence change ("Explore-Old"; blue), and exploratory rewards after the sequence change ("Explore-New"; red) in SNc (top row) and LC (bottom row) with a short (b) or long (c) time axis. **d.** Left: Schematic showing full trace (black), slowly-varying component (blue), and quickly-varying component (red). Right: PSTHs time-

locked to reward for SNc (top row) and LC (bottom row) using the slowly-varying components (left column, blue) or quickly-varying components (right column, red). **e.** PSTHs of activity when all sequences are rewarded with high (80%; blue) or low (20%; red) probability relative to effects seen in exploitative states (black) in SNc (top) or in LC (bottom). **f.** PSTHs of individual neurons that exhibit changes in sustained levels of activity during exploitative states (black) in SNc (top) and LC (bottom) time-locked to performance of the rewarded sequence. The mean activity of other neurons in the network during exploitative states (red) and the mean activity of all cells during exploratory states (blue) are shown for comparison. Insets: Proportion of the network exhibiting sustained activity during exploitative and exploratory states in SNc (top) and in LC (bottom). **g.** Individual neurons exhibit a range of correlations with RL estimates of action value (black), state value (red), or RPE (blue). **h.** Neurons with positive activity correlations (>0.1) with either action value (left), state value (middle), or RPE (right) do not exhibit different levels of sustained activity relative to cells with negative correlations (<-0.1). Error bars denote s.e.m.

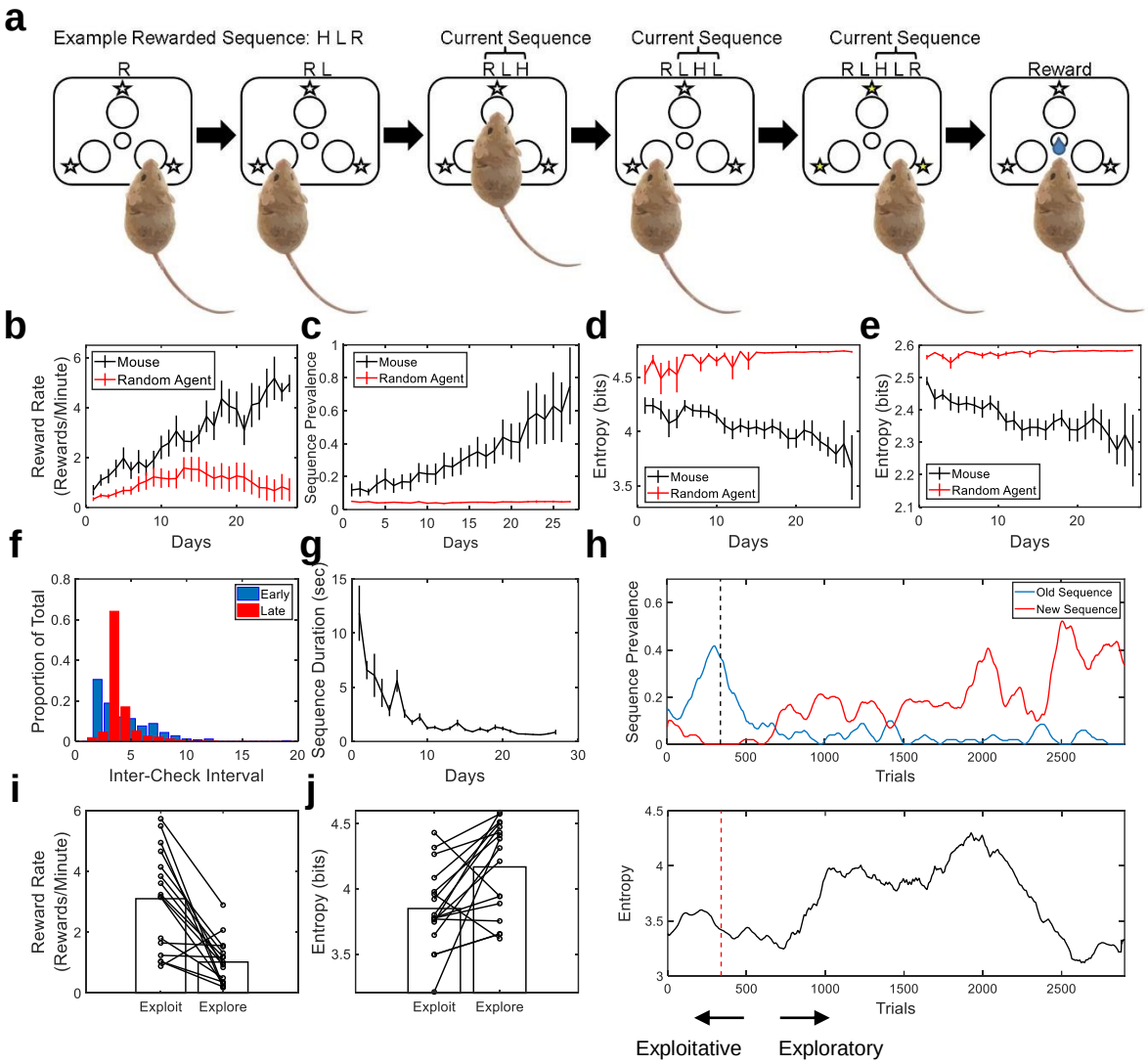
Figure 3. Activity accumulates positively in SNc and negatively in LC during exploitative action-reward bouts. a. Average occurrence of action-reward events time-locked to the occurrence of other action-reward events. There is no change in the mean temporal profile of action-reward occurrences between exploitative and exploratory states. **b.** Action-reward bouts are defined as clusters of action-reward pairs separated from each other by less than 10 seconds and separated from other action-reward pairs by more than 20 seconds. **c.** The raw number of action-reward bouts is increased in exploitative relative to exploratory states. Bottom: Example raster showing occurrences of action-reward pairs clustering more in bouts during exploitation relative to exploration. **d-g.** Mean baseline activity preceding action-reward events separated by position

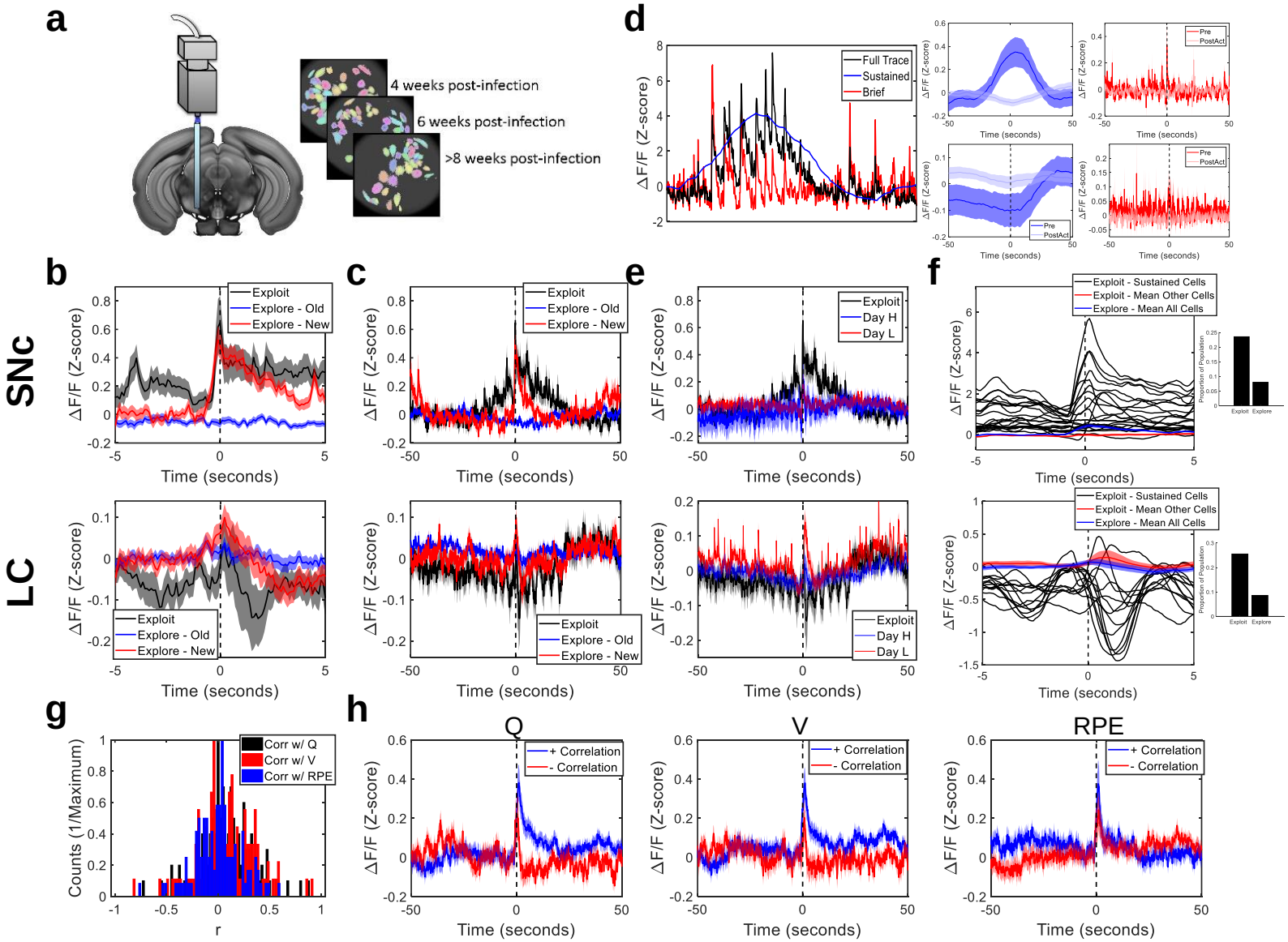
within action-reward bouts in SNc (top) and LC (bottom) during exploitative states (c), during exploratory states (d), on Day H (e), and on Day L (f), **h.** Neurons in SNc (top) and LC (bottom) whose activity is predictive (blue) and non-predictive (red) of reward bouts using a wiener filter. Error bars denote s.e.m.

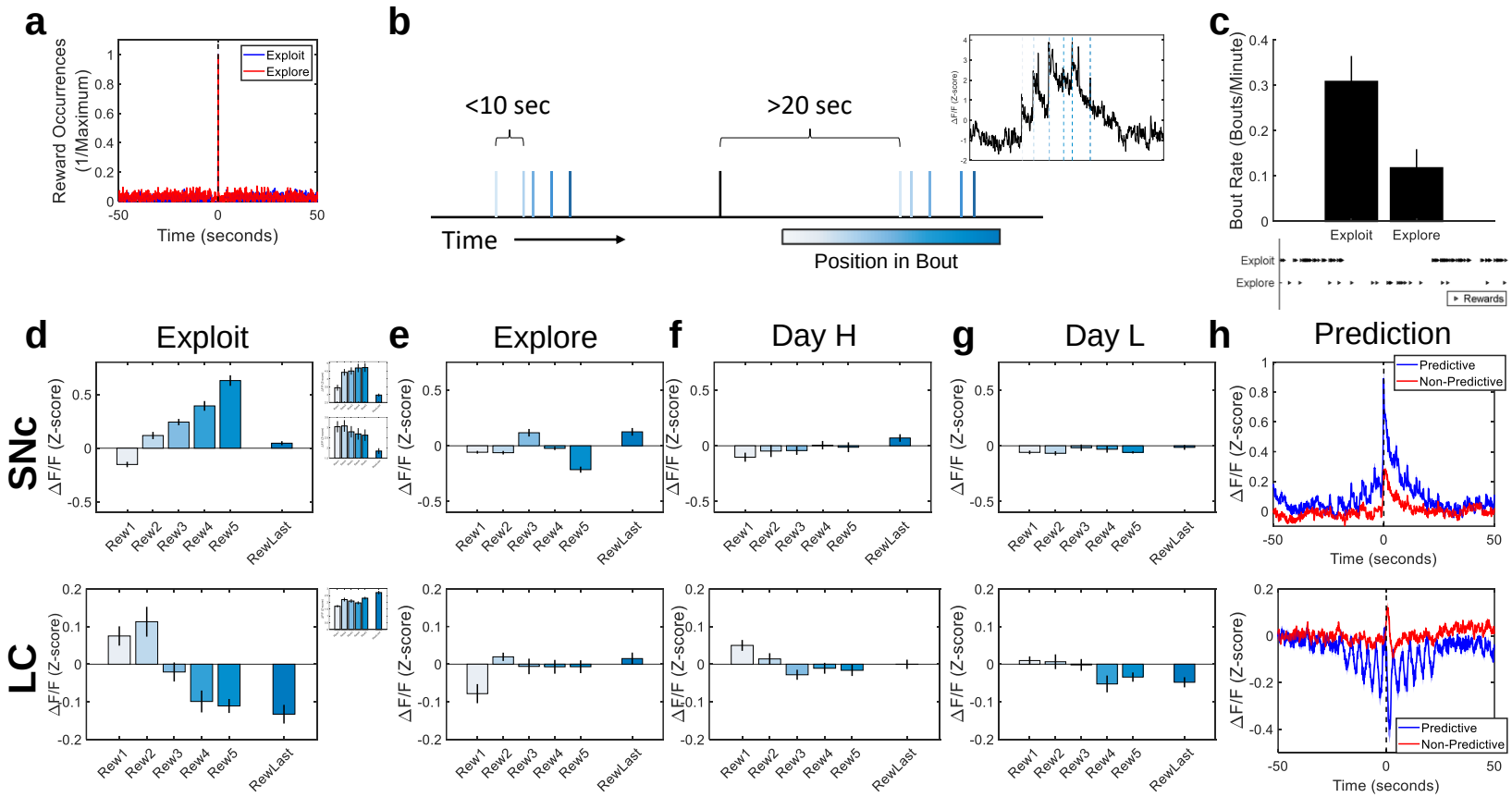
Figure 4. Extended response plateaus in SNc and depressions in LC produce sustained accumulations of activity. a,c. Mean durations of positive and negative transients in SNc (a), and in LC (c). Insets: Examples of extended response plateaus (a) and depressions (c). **b,d.** Cross-correlation histogram of neuronal activity time-locked to large fluorescence bursts in other cells during exploitative (blue) and exploratory (red) states in SNc (j) and LC (l). **e-f.** Schematics showing IRFs used to produce reward convolution traces for SNc (a) and LC (b). **g-j.** Reward convolution traces with typical IRFs in SNc (g) and LC (i), and with varying length IRFs in SNc (h) and LC (j). Inset (h): Reward convolution traces including hysteretic dynamics. **k-l.** Reward convolution model responses to Exploit (black), Explore-Old (blue), and Explore-New (red) with a range of IRFs in SNc (k) and LC (l). Error bars denote s.e.m.

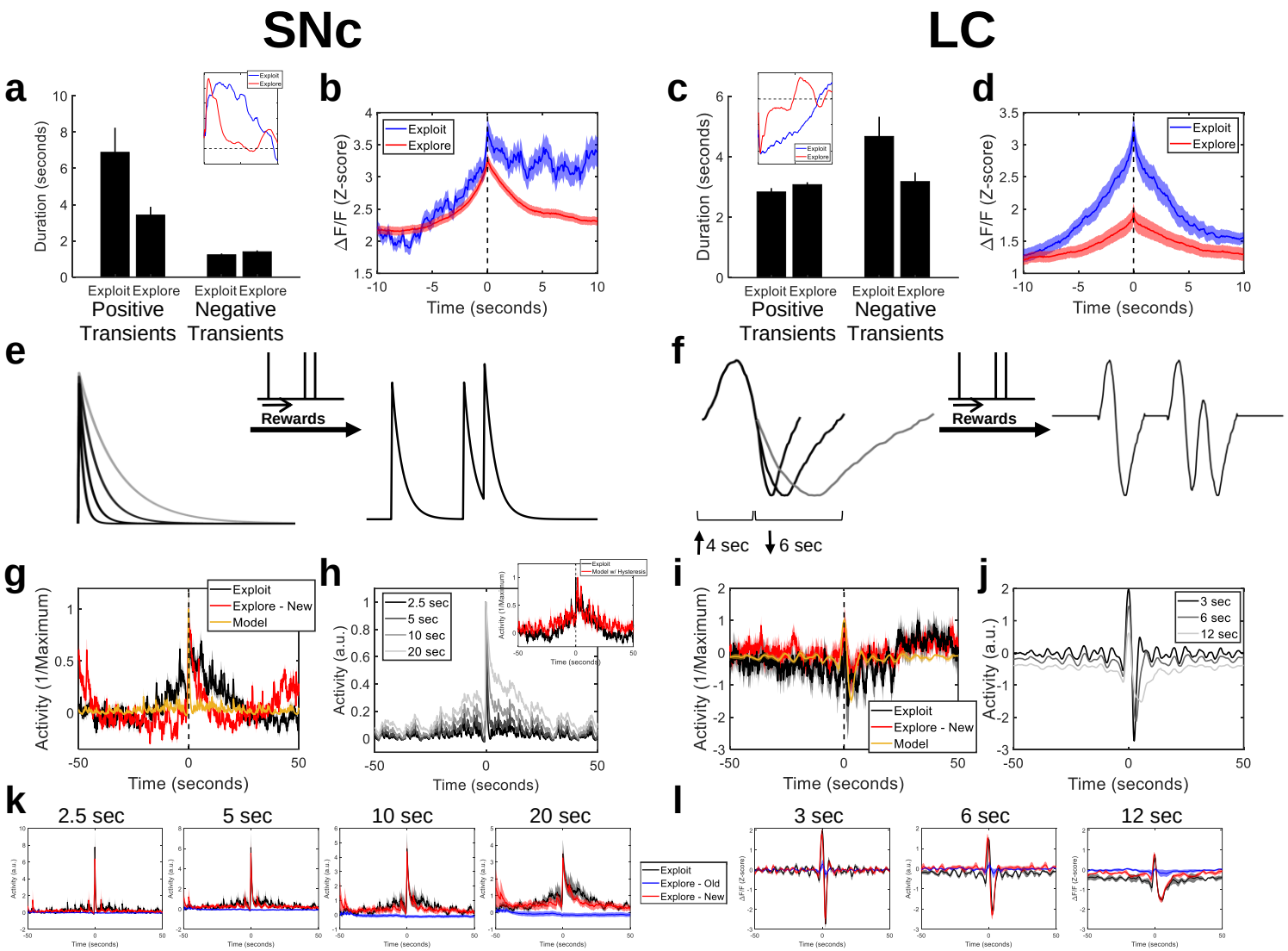
Figure 5. Increasing excitability in dopaminergic and noradrenergic neurons modulates exploitative and exploratory states. a. Experimental timeline. **b.** CNO injections result in an increase in the proportion of action-reward pairs that occur in bouts (left) and an increase in the number of action-reward pairs per bout (right) in animals expressing hM3Dq in SNc on Day H, but not in animals expressing mCherry in SNc or on other experimental days. Bottom: Example raster showing the timing and clustering of action-reward events following injections of VEH (top) or CNO (bottom). **c.** CNO injections result in an increase in transition entropy (left) and a decrease

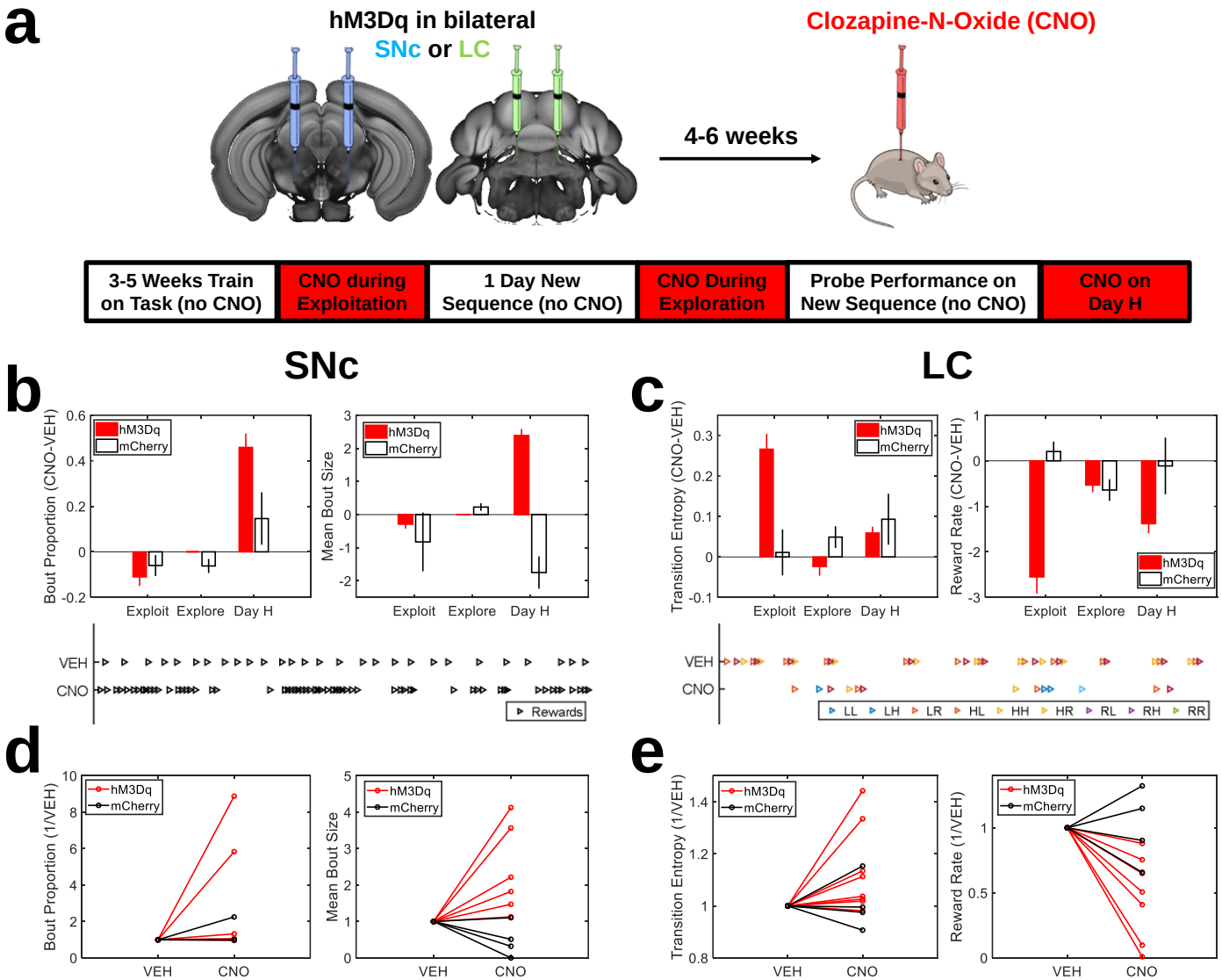
in reward rate (right) in animals expressing hM3Dq in LC during exploitative states, but not in animals expressing mCherry in LC or on other experimental days. Bottom: Example raster showing the selected transitions between nose pokes following injections of VEH (top) or CNO (bottom). Transitions are less stereotyped following CNO. **d.** Proportion of action-reward pairs that are in bouts (left column) and mean number of action-reward pairs per bout (right column) in animals expressing hM3Dq (red) or mCherry (black) in SNc following injections of either VEH or CNO, normalized to levels seen following VEH injections. **e.** Transition entropy (left column) and reward rate (right column) in all animals expressing hM3Dq (red) or mCherry (black) in LC following injections of either VEH or CNO, normalized to levels seen following VEH injections. Error bars denote s.e.m.


a
Example Rewarded Sequence: H L R
R
R L
Current Sequence
R L H
Current Sequence
R L H L
Current Sequence
R L H L R
Reward
b
Reward Rate (Rewards/Minute)
Mouse
Random Agent
Days
c
Sequence Prevalence
Mouse
Random Agent
Days
d
Entropy (bits)
Mouse
Random Agent
Days
e
Entropy (bits)
Mouse
Random Agent
Days
f
Proportion of Total
Early
Late
Inter-Check Interval
g
Sequence Duration (sec)
Days
h
Sequence Prevalence
Old Sequence
New Sequence
Trials
i
Reward Rate (Rewards/Minute)
Exploit
Explore
j
Entropy (bits)
Exploit
Explore
Entropy
Trials
Exploitative
Exploratory


a
4 weeks post-infection
6 weeks post-infection
>8 weeks post-infection
d
Full Trace
Sustained
Brief
ΔF/F (Z-score)
Time (seconds)
Pre
PostAct
SNc
LC
b
c
e
f
Exploit
Explore - Old
Explore - New
Day H
Day L
Exploit - Sustained Cells
Exploit - Mean Other Cells
Explore - Mean All Cells
Proportion of Population
Exploit
Explore
g
h
Counts (1/Maximum)
r
Corr w/ Q
Corr w/ V
Corr w/ RPE
Q
V
RPE
+ Correlation
- Correlation


a
Reward Occurrences (1/Maximum)
Exploit
Explore
Time (seconds)
b
<10 sec
>20 sec
Time
Position in Bout
c
Bout Rate (Bouts/Minute)
Exploit
Explore
Rewards
d
Exploit
e
Explore
f
Day H
g
Day L
h
Prediction
SNc
LC
ΔF/F (Z-score)
Predictive
Non-Predictive
Rew1
Rew2
Rew3
Rew4
Rew5
RewLast


SNc
LC
a
Duration (seconds)
Exploit Explore
Positive Transients
Exploit Explore
Negative Transients
b
ΔF/F (Z-score)
Exploit
Explore
Time (seconds)
c
Duration (seconds)
Exploit Explore
Positive Transients
Exploit Explore
Negative Transients
d
ΔF/F (Z-score)
Exploit
Explore
Time (seconds)
e
Rewards
f
Rewards
4 sec
6 sec
g
Activity (1/Maximum)
Exploit
Explore - New
Model
Time (seconds)
h
Activity (a.u.)
2.5 sec
5 sec
10 sec
20 sec
Exploit
Model w/ Hysteresis
Time (seconds)
i
Activity (1/Maximum)
Exploit
Explore - New
Model
Time (seconds)
j
Activity (a.u.)
3 sec
6 sec
12 sec
Time (seconds)
k
2.5 sec
5 sec
10 sec
20 sec
Exploit
Explore - Old
Explore - New
l
3 sec
6 sec
12 sec


a
hM3Dq in bilateral
SNc or LC
Clozapine-N-Oxide (CNO)
4-6 weeks
3-5 Weeks Train on Task (no CNO)
CNO during Exploitation
1 Day New Sequence (no CNO)
CNO During Exploration
Probe Performance on New Sequence (no CNO)
CNO on Day H
b
SNc
c
LC
Bout Proportion (CNO-VEH)
Mean Bout Size
Transition Entropy (CNO-VEH)
Reward Rate (CNO-VEH)
hM3Dq
mCherry
Exploit
Explore
Day H
VEH
CNO
Rewards
LL
LH
LR
HL
HH
HR
RL
RH
RR
d
e
Bout Proportion (1/VEH)
Mean Bout Size
Transition Entropy (1/VEH)
Reward Rate (1/VEH)

MATERIALS AND METHODS

Animals

All experiments were performed in compliance with the regulations of the Institutional Animal Care and Use Committee (IACUC) at Columbia University. A total of forty-eight mice (13 female, 35 male) of roughly 3 months of age were used for the experiments. Transgenic mice expressed Cre recombinase under the control of the tyrosine hydroxylase promoter (Tg(Th-cre)FI12Gsat/Mmucd) for targeting of dopaminergic and noradrenergic cells, or Cre recombinase under the control of the dopamine transporter promoter (B6.SJL-Slc6a3tm1.1(cre)Bkmn/J) for targeting of dopaminergic cells.

Virus Injections

Surgeries were performed under sterile conditions using isoflurane anesthesia (1-3%). Stereotactic coordinates relative to bregma were used to target the SNc (anteroposterior -3.16 mm, mediolateral ±1.4 mm, dorsoventral -4.2 mm) and stereotactic coordinates relative to lambda were used to target the LC (anteroposterior -0.8 mm, mediolateral ±0.8 mm, dorsoventral -3.2 mm). For imaging experiments, animals were injected unilaterally with 500 µL of AAV5.CAG.Flex.GCaMP6f.WPRE.SV40 (University of Pennsylvania Vector Core) into the right SNc or LC. For chemogenetic experiments, experimental animals were injected bilaterally with 500 µL of AAV5-hSyn-DIO-hM3D(Gq)-mCherry (Addgene plasmid #44361), while control animals were injected bilaterally with 500 µL of AAV5-hSyn-DIO-mCherry (Addgene plasmid #50459). All injections were performed using a Nanoject II Injector (Drummond Scientific, Broomall, PA, USA) at a rate of 4.6 nL every 5 seconds. Injection pipettes were left in place for 10 minutes post-injection to allow for virus absorption, and incisions were closed with Vetbond

tissue adhesive (3M, Maplewood, MN, USA) for chemogenetic experiments in which no lens was implanted. Animals were given a minimum of 5 days to recover from surgery before behavioral training.

Chronic Lens Implantation

For imaging experiments, virus injections were followed by implantation of a gradient index (GRIN) lens (Inscopix, Inc., Palo Alto, CA, USA) into the SNc or LC. Overlying tissue was first removed by insertion of a 30-gauge blunt needle to the target site, with care taken to minimize damage. GRIN lenses were then implanted unilaterally and secured to the skull using dental acrylic (Lang Dental, Wheeling, IL, USA). 2-3 weeks were allowed for virus expression before attachment of microendoscope baseplates (Inscopix, Inc.) to the dental acrylic at the correct focal plane for imaging.

Chemogenetics

For chemogenetic experiments, mice were briefly anesthetized with 1-3% isoflurane and injected intraperitoneally with 5 mg/kg clozapine-N-oxide (CNO) before behavioral sessions. Mice were given 15 minutes following CNO injection to allow for the CNO to take effect and for the isoflurane effects to subside.

Behavioral Task

Animals were trained in custom-made operant boxes (5 in x 6 in) controlled by a python-based framework (PyControl, https://pycontrol.readthedocs.io) that supplies all cues and rewards, as well as recording all behavioral timestamps. Behavior was also monitored with overhead

cameras (Flea3, Point Grey Research, Richmond, Canada) recording at 15-30 frames per second. Operant boxes were placed inside sound attenuating chambers during training. Timestamps from the behavioral task were synchronized with calcium imaging data using TTL pulses sent from the behavioral chambers to the Inscopix data acquisition system via a BNC cable.

Operant chambers contained three equidistant nose poke ports surrounding a central reward port. Mice had to discover a rewarded sequence of three pokes in a specific order with no intervening pokes. Importantly, the task contains no trial structure and few cues, ensuring that mice actively explore the environment to discover what is rewarding. When a correct sequence was performed, water rewards of 5-15 µL were supplied through the opening of a solenoid. Throughout the paper, when trials are mentioned, they refer to the performance of single nose pokes.

Mice were initially pre-trained in a setting in which any possible three-poke sequence that includes all three nose poke ports was rewarded. Following roughly one week of pre-training, mice were exposed to the full task, in which only one target sequence was rewarded. Once mice achieved proficiency on a particular target sequence, the rewarded sequence was changed. For experiments on Day H, all three-poke sequences were rewarded with 80% probability. For experiments on Day L, all three-poke sequences were rewarded with 20% probability. In all cases, rewards could not be cached and had to be consumed prior to earning further rewards.

During calcium imaging experiments, fluorescence images were acquired at a frame rate of 10 hz.

Data Analysis

Analyses were performed in Matlab (Mathworks, Natick, MA) with custom-written routines. Behavioral data were sampled in 1 ms bins. For sliding window analyses of behavioral

data, a window size of 100 trials with a step size of 5 trials was used. For each behavioral session, histograms were created for the empirical probability of performing each sequence ("Sequence Entropy"), as well as for the empirical probability of transitions between nose pokes ("Transition Entropy"), and entropy was calculated as:

$$\mathrm{H}(X) = -\sum_{i=1}^{n} P(x_i) \log_2 P(x_i)$$

Corrections for finite sample sizes[13,42] were tested by sampling from a known distribution with a structure similar to that seen in our behavioral data, and these corrections were found to be less accurate than the above formula in measuring the entropy of the parent distribution. These corrections were therefore not used in subsequent analyses.

A basic reinforcement learning (RL) model was also applied to the behavioral data. Expected action values, Q, were updated on each trial according to:

$$Q_{t+1}(c_t) = Q_t(c_t) + \alpha\delta_t$$

where c_t is the choice on trial t, $\delta_t$is the reward prediction error on trial t, and α is the learning rate of the model. Expected action values were related to choices by the following equation:

$$p\big(c_t(a)\big) = \frac{e^{\beta Q_t(a)}}{\sum_{b=1}^{n} e^{\beta Q_t(b)}}$$

where β is the inverse temperature parameter. Finally, expected state values, V, were estimated as the sum of all current action values in that state weighted by their probability of occurrence:

$$V_t(s_t) = \sum_{i=1}^{n} Q_t(c_t)p(c_t)$$

The learning rate and inverse temperature were fit using maximum likelihood estimation.

Occurrences of reward bouts were predicted from neural data using a wiener filter. Five lags were used occurring every 500 milliseconds starting 2 seconds before bout start and ending

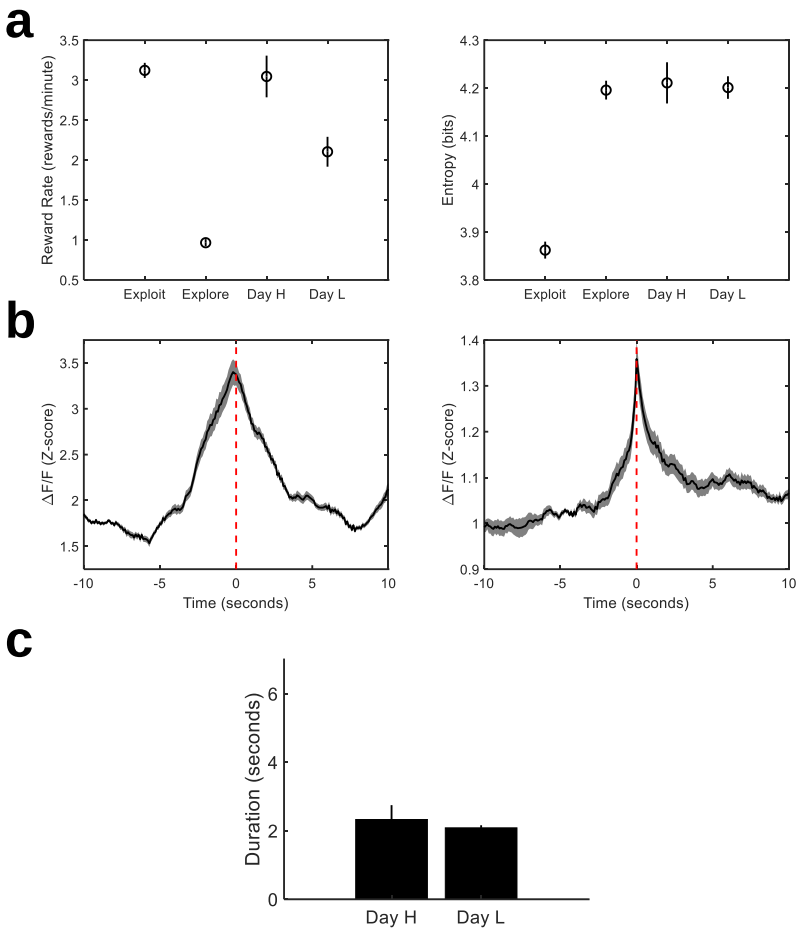
at bout start. Prediction was done for each cell individually to assess their contribution. Prediction performance was assessed using the area-under-the-curve (AUC) from a receiver operating characteristic (ROC) curve. Results were also compared to results obtained when the behavioral category labels were shuffled.

For reward convolution models, pure exponentials of varying lengths were used to model the SNc impulse response function. For the LC impulse response function, the average response from all LC cells to unexpected rewards was smoothed by a 1 second moving average. To model hysteretic network dynamics, multiple convolution traces were created for each animal with the addition of a random temporal jitter in the response to reward for each trace that preferentially shifts responses positively in time by a random fraction of a maximum of 5 seconds.

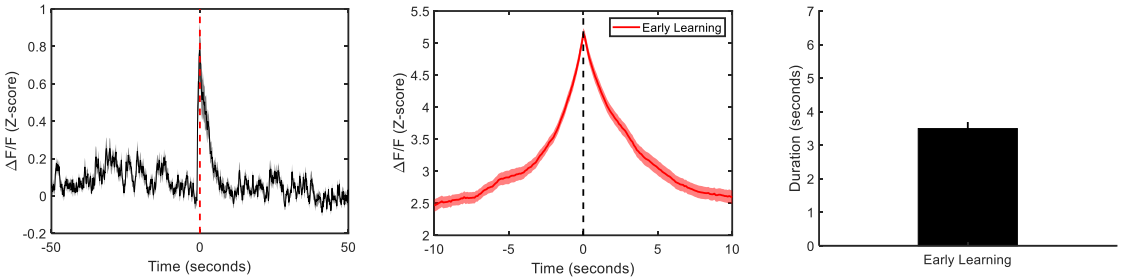
To quantify response plateau and depression durations, positive and negative threshold crossings (3SD) were located. A 5 second window before threshold crossing was defined as baseline activity, and a 5 second window after threshold crossing was then advanced until the average activity in this window matched the average activity in the baseline window. The number of timepoints by which the second window had to be advanced was defined as the response plateau or depression duration.

Calcium imaging data were first preprocessed using Mosaic (Inscopix, Inc.) to apply 4x spatial downsampling and motion correction. Constrained non-negative matrix factorization (CNMF-E)[43-44] was then applied for denoising and demixing of the data. The footprints and activity profiles of all putative neurons were inspected manually before inclusion in the dataset. For the extraction of quickly-varying components of fluorescence signals, dF/F was calculated on these traces with a sliding window of 5 seconds. For the extraction of slowly-varying (tonic) components of the fluorescence signals, traces were smoothed with a moving average of 60 seconds.

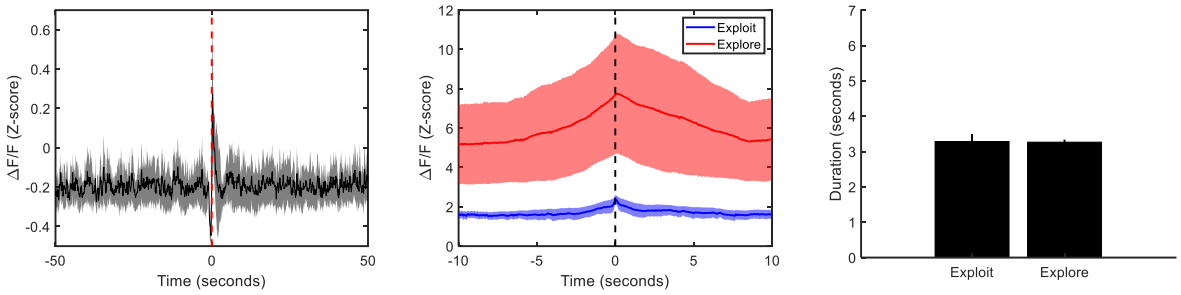
SUPPLEMENTARY FIGURES



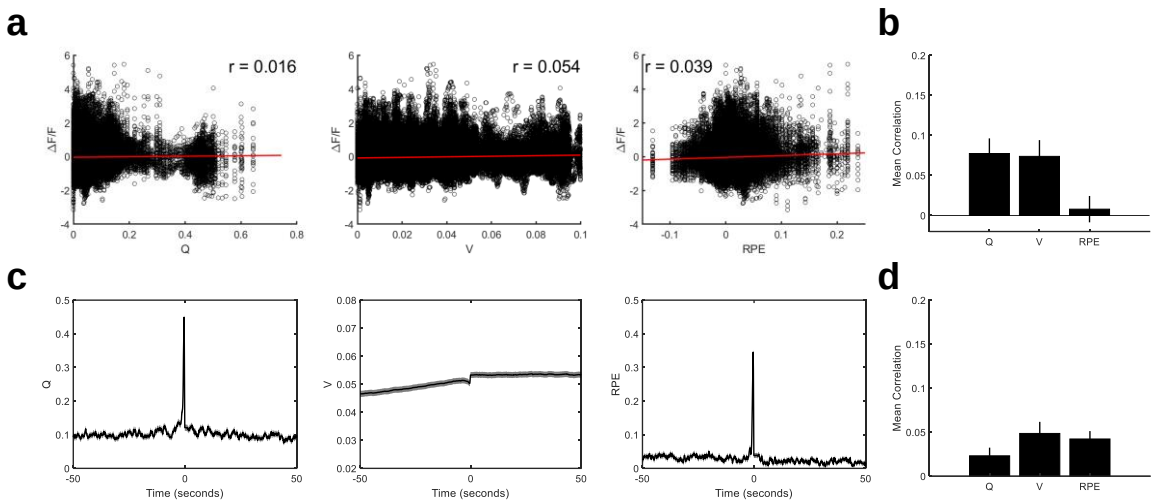
Supplementary Figure 1. Behavioral and neural metrics on Day H and Day L. a. Reward rate (left) and response entropy (right) during exploitation, exploration, on Day H, and on Day L. **b.** Cross-correlation histograms of activity in SNc time-locked to large fluorescence bursts in other cells in the network on Day H (left) and Day L (right). **c.** Duration of plateaus on Day H and Day L.



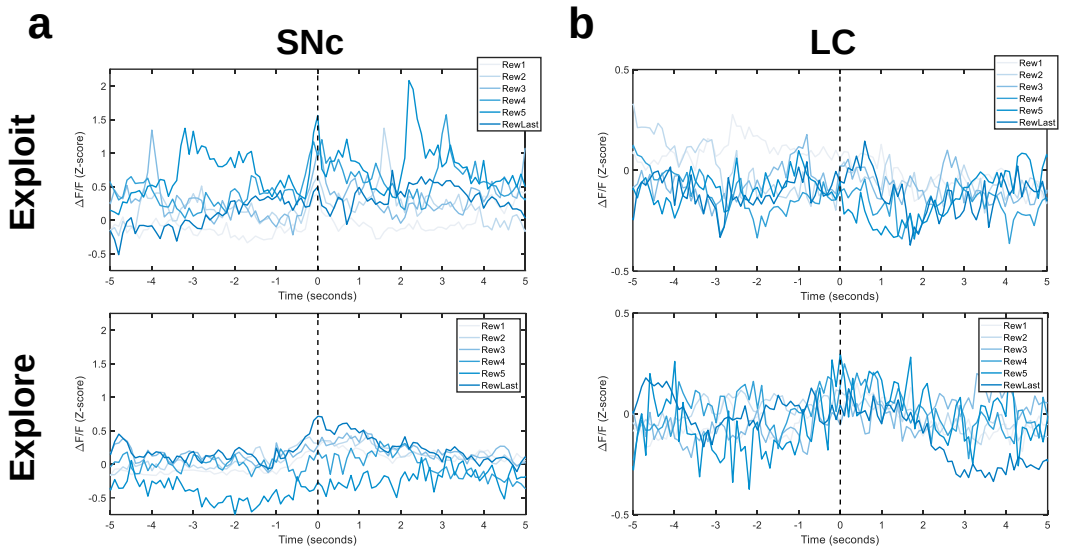
Supplementary Figure 2. SNc effects are not present early in training. PSTH of SNc activity time-locked to exploitative rewards (Left), cross-correlation histogram (center), and plateau duration (right) early in training.



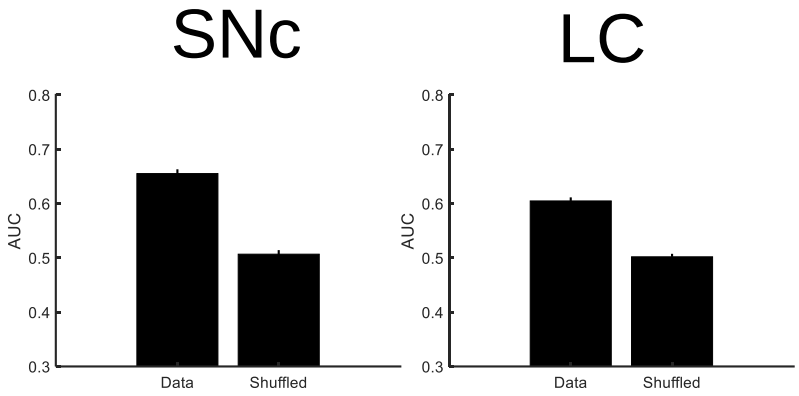
Supplementary Figure 3. Effects observed in SNc are not present in VTA. PSTH of VTA activity time-locked to rewards (Left), cross-correlation histogram (center), and plateau duration (right) during exploitative states.



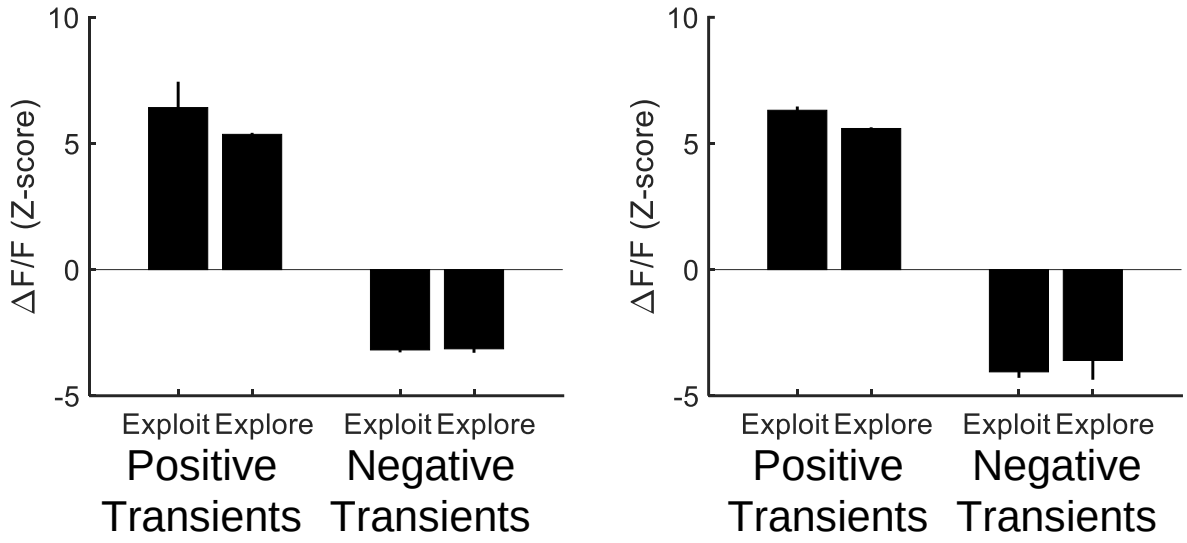
Supplementary Figure 4. SNc activity is not correlated with basic RL parameters. a. Scatter plots showing mean neuronal activity within 2 seconds of an action versus that trial's current estimate of action value (left), state value (center), and reward prediction error (right). **b.** Mean correlations of peri-event neuronal activity in individual cells with action-by-action estimates of action value, state value, and reward prediction error. **c.** PSTHs time-locked to reward using action-by-action estimates of current action value (Q, left), state value (V, center), or reward prediction error (RPE, right). **d.** Mean correlation of full session time courses of action value, state value, and reward prediction error with smoothed SNc neuronal activity.



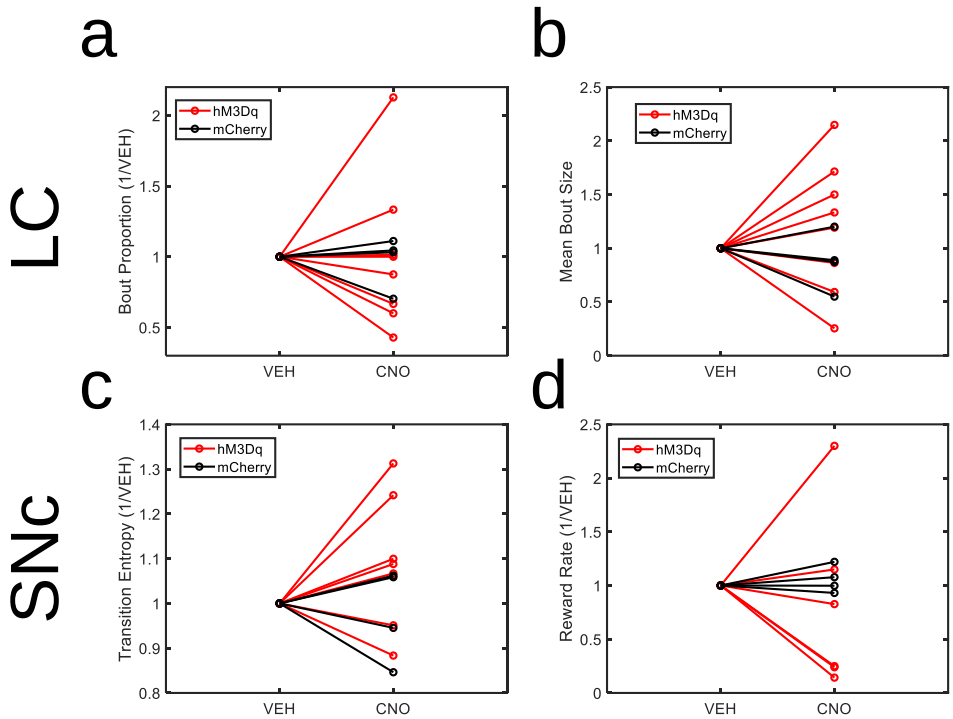
Supplementary Figure 5. Activity accumulations within reward bouts. a. PSTHs of SNc activity time-locked to reward during exploitative (left) and exploratory (right) states. **b.** PSTHs of LC activity time-locked to reward during exploitative (left) and exploratory (right) states.



Supplementary Figure 6. Wiener filter prediction performance. Area-under-the-curve assessment of wiener filter performance using activity from individual SNc (left) or LC (right) cells relative to prediction performance in cases in which the category labels were shuffled.



Supplementary Figure 7. Positive and negative transient amplitude does not change across behavioral states. Amplitude of positive and negative transients in exploitative and exploratory behavioral states in SNc (left) and LC (right). The amplitude of transients does not change and cannot account for the observed changes in sustained activity.



Supplementary Figure 8. Chemogenetic effects are specific to dopaminergic and noradrenergic systems. a. The proportion of rewards that are in reward bouts does not change when animals expressing hM3Dq in LC are given CNO versus VEH. **b.** The mean number of action-reward pairs in action-reward bouts does not change when animals expressing hM3Dq in LC are given CNO versus VEH. **c.** The transition entropy does not change when animals expressing hM3Dq in SNc are given CNO versus VEH. **d.** The overall reward rate does not change when animals expressing hM3Dq in SNc are given CNO versus VEH.