NEUROSCIENCE

Dichotomous dopaminergic and noradrenergic neural states mediate distinct aspects of exploitative behavioral states

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

The balance between exploiting known actions and exploring alternatives is critical for survival and hypothesized to rely on shifts in neuromodulation. We developed a behavioral paradigm to capture exploitative and exploratory states and imaged calcium dynamics in genetically identified dopaminergic and noradrenergic neurons. During exploitative states, characterized by motivated repetition of the same action choice, dopamine neurons in SNc encoding movement vigor showed sustained elevation of basal activity that lasted many seconds. This sustained activity emerged from longer positive responses, which accumulated during exploitative action-reward bouts, and hysteretic dynamics. Conversely, noradrenergic neurons in LC showed sustained inhibition of basal activity due to the accumulation of longer negative responses in LC. Chemogenetic manipulation of these sustained dynamics revealed that dopaminergic activity mediates action drive, whereas noradrenergic activity modulates choice diversity. These data uncover the emergence of sustained neural states in dopaminergic and noradrenergic networks that mediate dissociable aspects of exploitative bouts.

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 known actions (low choice entropy) and exploring alternative ones (high choice entropy) has been proposed to rely on midbrain dopaminergic neurons of the substantia nigra pars compacta (SNc) (*1*–*3*) and noradrenergic neurons of the locus coeruleus (LC) (*4*–*5*). Deficits in choice reversal learning (*6*) and attentional set-shifting (*7*) have been demonstrated following dopamine (DA) depletion, and computational modeling work has predicted a central role for DA signaling in modifying action selection probabilities (*8*). Similarly, recent work suggests that levels of norepinephrine (NE), specifically the noradrenergic projections to prefrontal cortices, modulate levels of stochastic responding in rodents (*9*–*10*). Both dopaminergic and noradrenergic systems therefore appear to play a central role in motivating and structuring adaptive behavior.

Although past work has studied isolated exploitative and exploratory choices (*1*–*2*, *11*–*15*), most 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. While animals often make lone exploratory actions during exploitative states in stable environments, the current work focuses on longer-term exploratory states, similar to scenarios in which animals must abandon an overforaged location in search of a new, more rewarding location (*16*). We will refer to these longer-term states here as exploitative and exploratory behavioral states to distinguish them from more isolated action choices. These exploitative states, characterized by periods of engaged and motivated performance of the same well-learned action to achieve a desired outcome, might be similar to what is colloquially referred to as "being in the zone" or the "hot-hand effect" (*17*–*18*). However, although both the motivational and repeated choice aspects of exploitative states go hand in hand, it is unclear whether they are mediated by the same neural substrates.

We developed a novel behavioral task in mice that probes action selection among many possible actions over long time scales and allows us to bias behavior toward exploitative or exploratory states using environmental changes. This paradigm permitted us to study the behavior of animals away from ceiling or floor performance and hence to study the emergence of bouts of exploitative choices, i.e., periods of motivated performance of the same actions. We imaged the activity of populations of individual DA neurons of the SNc and noradrenergic neurons of the LC and found notable changes in sustained dopaminergic and noradrenergic activity that cumulatively emerged when animals were in exploitative behavioral states. These exploitative states were marked by lengthened response plateaus and hysteretic network dynamics in SNc neurons, as well as lengthened response depressions in LC neurons. The sustained activity changes of SNc, but not LC, neurons were related to the vigor or motivation of the behavior. Last, we induced sustained changes in the excitability of SNc dopaminergic neurons and LC noradrenergic neurons and found that these systems subserve dissociable aspects of action motivation and selection, with SNc mediating the motivation to engage in action bouts and LC mediating choice entropy. These data reveal that these two major neuromodulatory systems display sustained neural states that mediate different aspects of exploitative behavioral states.

RESULTS

Sustained dopaminergic and noradrenergic modulations in a novel task probing exploitative and exploratory behavioral states

To develop a framework for studying exploitative and exploratory states in mice, we created a nose-poke sequence task in which mice

[1] Zuckerman Mind Brain Behavior Institute, Columbia University, New York, NY, USA. [2] Champalimaud Neuroscience Programme, Champalimaud Centre for the Unknown, Lisbon, Portugal.
*Corresponding author. Email: rc3031@columbia.edu

can choose between many possible actions. Mice were placed in an operant chamber with three equidistant nose pokes (Fig. 1A). A sequence of three pokes in a specific order was rewarded. Mice were given no trial structure 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 is a large action space with 27 possible sequences, providing a broad distribution of potentially selectable actions, and we used the entropy of this distribution to assess levels of exploitative and exploratory behavior.

Performance improved significantly over training, as seen both in an increase in reward rate [$P = 1.08 \times 10^{-7}$, $t(16) = 8.23$, and $N = 17$ animals; Fig. 1B] and an increase in the proportion of pokes that compose the rewarded sequence relative to total pokes [$P = 1.26 \times 10^{-4}$ and $t(16) = 4.79$; 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. 1, B to E, gray lines). Chance level of reward rate was, on average, 0.813 rewards/min, and after training, animals performed well above chance level for all behavioral measures ($P < 0.001$). After training, mice were not only well above floor performance but also below ceiling performance, allowing us to study the transitions between periods of exploitative and exploratory responding within these broader behavioral states. Over the course of training, we observed a decrease in the entropy of the animals' selected sequences [$P = 0.019$ and $t(16) = 2.73$; Fig. 1D] and a decrease in the entropy of the animals' selected transitions between nose pokes [$P = 0.0015$ and $t(16) = 3.71$; 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 structure behavior into groups of three nose pokes in late learning (Fig. 1F). The time that mice took to perform the rewarded sequence also decreased significantly with training [$P = 0.0079$ and $t(16) = 2.97$; Fig. 1G].

When the animals reached plateau performance for a particular target sequence, we changed the target sequence to be rewarded. We always implemented the sequence change in the middle of a single behavioral session, so as to have both exploitative and exploratory phases in the same session. Across all animals, we observed a significant decrease in performance [$P = 8.68 \times 10^{-5}$ and $t(16) = 5.20$; Fig. 1H] and an increase in entropy [$P = 0.004$ and $t(16) = 3.36$; Fig. 1I] immediately following the rule change, suggesting that changing the contingency between action and reward successfully drove animals into a more exploratory state, where they explore task parameters to find the new rule (figs. S1 and S2). Although learning and exploration are intricately related (*19*), this rule change occurred after animals had ample experience with all task-related actions, and therefore, this state reflects an exploration of known actions rather than learning a specific action de novo. The relatively large size of the action space that the mice must explore in this task following a reversal provides an extended exploratory period for analysis that spans multiple behavioral sessions (fig. S2) before the new rewarded sequence is found and performed consistently.

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 following injection of viruses carrying cre-dependent GCaMP6f in TH-cre mice (Fig. 1J and fig. S3). We imaged a total of 121 SNc cells (7 animals) and 61 LC cells (10 animals) during task performance. Constrained nonnegative matrix factorization (CNMF-E) was used to extract regions of interest, fluorescence traces, and inferred spiking activity (*20*–*22*). We first examined phasic bursting in the mean population responses before and after a change in the rewarded sequence (Fig. 1K), 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 peri-stimulus time histograms time-locked to the final nose poke of the target sequence (simultaneous with the onset of reward cue). 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 "explore-old," representing perseverative errors when mice performed the previously rewarded action that was no longer rewarded, and "explore-new," representing trials when mice performed the newly rewarded action (both time-locked to the end of performance of the relevant action). Sequence changes occurred within a single behavioral session, ensuring that levels of fluorophore expression and bleaching, as well as satiety, were comparable across comparisons. In both regions, we observed qualitatively similar phasic responses to rewards during exploitative and exploratory epochs. However, these phasic responses arose from different baseline levels of activity, with SNc baseline activity enhanced, and LC baseline activity reduced, during exploitation ($P < 0.05$ and $N = 121$ SNc cells, 61 LC cells; Fig. 1K). In SNc, exploitative and exploratory rewards resulted in comparable peak magnitudes, despite the change in baseline activity, consistent with recent reports suggesting that reward expectation is marked by increased baseline activity rather than decreased peak amplitude (*23*). Expanding the time axis, we found that these baseline changes developed slowly across multiple trials, lasting for roughly 30 s surrounding exploitative rewards (Fig. 1L). This is a very large temporal window for analysis that contains multiple other actions and rewards, the responses to which are reflected in the group average. Nevertheless, there is a notable difference in neural activity between exploitative and exploratory states, which could represent a difference in the summation of neural responses across these states, a difference in the temporal structure of behavior across these states, or a combination of the two.

Therefore, we next asked whether these sustained activity changes were due simply to differences in reward rate during exploitation and exploration. We ran animals on versions of the task in which all possible three-poke sequences were 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 ($P < 0.05$; Fig. 1M). The reward rate on day H was comparable to that during exploitation but the entropy was significantly higher (fig. S4). This suggests that basic reward rate cannot account for the sustained baseline effects, although subtler changes to the high-level structure of rewards or to the neural responses to rewards could still be involved. We did not observe sustained changes in baseline activity in SNc or LC when we aligned our data to the last nose poke before a random check of the reward port that does not lead to reward, further suggesting that the observed sustained changes are related to the high-level structure of exploitative behavior rather than single actions or action-by-action estimates of expected value (fig. S5). We also did not observe sustained changes in baseline activity in early learning (fig. S6) or in dopaminergic neurons of

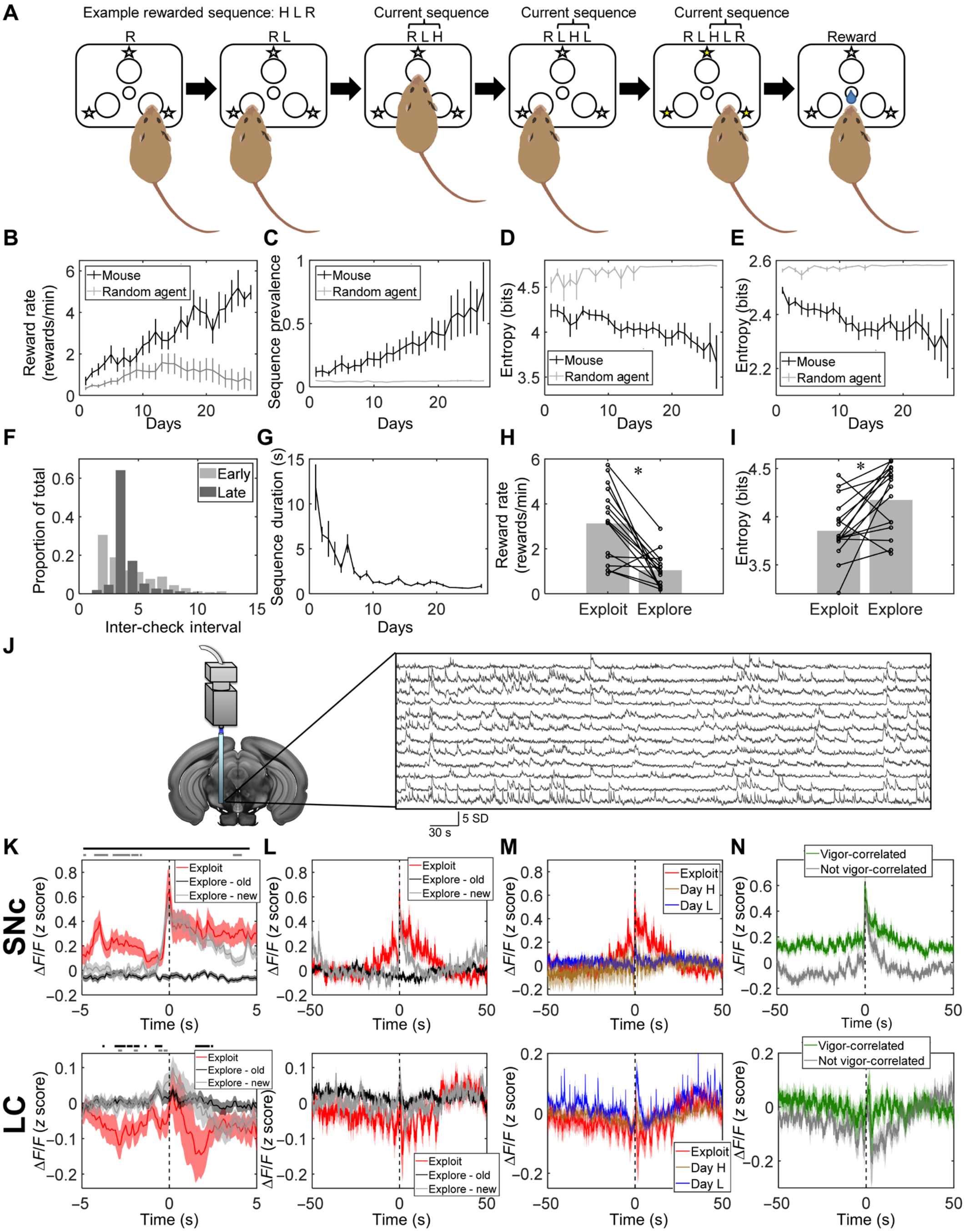

Fig. 1. Sustained dopaminergic and noradrenergic activity modulations in a novel task for probing action motivation and selection. (**A**) Task schematic. There are three nose-poke ports and mice must find a rewarded sequence of three nose pokes in order. A moving buffer of the last three nose pokes is monitored for the rewarded sequence. With training, (**B**) mean reward rate increases, (**C**) mean prevalence of the rewarded sequence increases, (**D**) mean entropy of the selected actions decreases, and (**E**) mean entropy of the transitions between pokes decreases. Gray lines indicate estimates of chance performance. (**F**) Histogram of the number of pokes at response ports between checks for reward at the reward port ("inter-check interval") in an example mouse in early (light gray) and late (dark gray) learning. (**G**) The average time to complete the rewarded sequence decreases with training. (**H**) Reward rate decreases and (**I**) entropy of selected sequences increases following a change in the rewarded sequence. (**J**) Schematic of endoscope imaging. Raw fluorescence traces from a random subset of the network are shown at right. (**K** and **L**) Mean population response time-locked to exploitative rewards before the sequence change ("exploit;" red), perseverative errors after the sequence change ("explore-old;" black), and exploratory rewards after the sequence change ("explore-new;" gray) in SNc (top) and LC (bottom) with a regular (K) or greatly extended (L) time axis. Bars above axes in (K) designate time points in which mean activity during exploit was significantly greater than explore-old (black) or explore-new (gray) at $P < 0.05$ or lower. (**M**) Mean population activity when all sequences are rewarded with high (80%; brown) or low (20%; blue) probability relative to effects seen in exploitative states (red) in SNc (top) or in LC (bottom). (**N**) Mean activity of vigor-related neurons (green) and non–vigor-related neurons (gray) time-locked to reward achievement during exploitative states. Error bars denote SEM. *, significant at $P < 0.05$ or lower. ns, $P > 0.05$.

the ventral tegmental area (VTA; fig. S7). This sustained effect was apparent not only in the population mean but also in individual neurons, and we found that roughly 20 to 25% of the recorded populations in SNc and LC exhibited these sustained changes in baseline activity during exploitation relative to exploration (fig. S8). We also found, as expected, that activity in a subset of the recorded cells was significantly correlated with estimates of action value (Q), state value (V), and reward prediction error (RPE) in a basic reinforcement learning (RL) model (see Materials and Methods), and these RL-correlated cells exhibited classical RPE responses (*24*) to exploitative or exploratory rewards (fig. S9). However, these cells did not exhibit the shifts in baseline activity that we observed in the network as a whole, suggesting that this subpopulation was not driving the sustained effect. In addition, we found no strong relationship between the cells that we classified as exhibiting sustained effects (in fig. S8) 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), again suggesting that populations correlated with RL parameters are independent from those exhibiting changes in sustained activity.

Therefore, we examined whether neurons exhibiting baseline shifts corresponded to subpopulations previously found in SNc related to movement initiation and vigor, which are distinct from those responding to reward (*25*). We found that the onset of neuronal responses in both regions aligned more closely with movement initiation than with reward achievement, while the response peak aligned more closely with reward achievement (fig. S10), consistent with past work demonstrating ramps in dopaminergic activity based on proximity to reward (*26–28*) and suggesting an association between activity in these neurons and the drive to act. When we then separated neurons based on the correlation of their activity with movement vigor (see Materials and Methods), we found that vigor-correlated SNc neurons had significantly higher baseline activity during exploitation than other recorded neurons ($P < 0.001$; Fig. 1N). The decreases in sustained activity that we observed in LC neurons during exploitation were not seen in vigor-correlated LC subpopulations (whose baseline activity remained at normal levels), but we instead observed these decreases in non–vigor-correlated LC subpopulations ($P < 0.05$; Fig. 1N). These data suggest that the observed baseline shifts in SNc occur preferentially in neurons whose activity reflects the motivation to perform actions but not in LC, which has previously been implicated in action choice (*4*, *9–10*).

Sustained activity cumulatively emerges in dopaminergic and noradrenergic networks during exploitative action-reward bouts

Although we found that the sustained activity changes were not due to average reward rate (Fig. 1M), we wondered whether the reward structure changed following the sequence change. Given that the SNc neurons that showed sustained activity seemed to be related to the motivation to perform an action, we investigated whether there were periods of repetition of the actions that lead to reward (similar to the hot-hand effect) during exploitative states. We therefore defined “target action-reward bouts” (hereafter “action-reward bouts”) as clusters of action-reward pairs (low choice entropy) that are separated from each other by less than 10 s and separated from other action-reward pairs by more than 20 s, and we then analyzed activity based on the action-reward pair’s position within a bout (Fig. 2A and fig. S11A). We found that the rate of occurrence of action-reward bouts increased significantly in exploitative relative to exploratory states [P = 0.013 and $t(8)$ = 3.17; Fig. 2B], suggesting that action-reward bouts could be an important characteristic of the transitions between these behavioral states. In addition, animals responded significantly faster during action-reward bouts relative to other times [$P = 5.46 \times 10^{-4}$ and $t(4103)$ = 3.46; fig. S11B], suggesting that these bouts represent periods of enhanced vigor and motivation to act. We therefore investigated neural responses throughout these bouts during exploitation and we found that baseline activity preceding action-reward events increased over the course of a bout in SNc [$P = 7.11 \times 10^{-4}/r = 0.992$ and $P = 2.55 \times 10^{-5}/F(4,238) = 6.97$; Fig. 2C, top], while LC baseline activity decreased consistently over the course of the bout [$P = 0.028/r = -0.917$ and $P = 0.026/F(4,230) = 2.82$; Fig. 2C, bottom]. This response pattern was also apparent in individual fluorescence traces during individual reward bouts (Fig. 2A, right), as well as when the analysis was performed using inferred spiking activity [SNc, $P = 0.006/r = 0.97$ and $P = 0.016/F(4,238) = 3.21$; LC, $P = 0.015/r = -0.95$ and $P = 0.006/F(4,230) = 3.64$; fig. S11C]. These patterns were not present during exploration [SNc, $P = 0.554/r = -0.37$ and $P = 0.009/F(4) = 3.46$; LC, $P = 0.39/r = 0.50$ and $P = 0.738/F(4) = 0.5$; Fig. 2D], despite there being no difference in the interval between rewards in action-reward bouts in exploitation relative to exploration (fig. S11D). Furthermore, these patterns were also not present on day H [SNc, $P = 0.007/r = 0.968$ and $P = 0.30/F(4,90) = 1.23$; LC, $P = 0.006/r = 0.97$ and $P = 0.26/F(4,255) = 1.32$; Fig. 2E] or on day L [SNc, $P = 0.67/r = 0.263$ and $P = 0.92/F(4,322) = 0.24$; LC, $P = 0.07/r = -0.85$ and $P = 0.746/F(4,385) = 0.49$; Fig. 2F]. Bouts of reward are common on day H; however, they do not correspond to repetition of the same target action on day H but rather to performance of bouts of different actions. Therefore, the hot-hand effect, or motivated repetition of the same action, should be minimal during this manipulation in natural settings, as many actions lead to reward.

We next asked whether the baseline activity profile we observed during action-reward bouts was sufficiently characteristic to enable prediction of bout occurrences based on neural activity. We therefore trained a Wiener filter to discriminate between action-reward pairs that occurred within bouts and action-reward pairs that occurred outside of bouts using baseline activity preceding those action-reward pairs, and we found that discrimination was significantly better than chance [SNc, $P = 1.27 \times 10^{-11}$ and $t(234) = 7.13$; LC, $P = 4.10 \times 10^{-7}$ and $t(110) = 5.39$; Fig. 2G, right]. Furthermore, if we considered only neurons whose activity was most predictive of action-reward bouts, then 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 ($P < 0.001$; Fig. 2G). Together, these data suggest that sustained activity accumulates positively in SNc and negatively in LC as animals perform bouts of the same target action sequence in exploitative, but not exploratory, states. Exploitative states are therefore marked by both a change in the structure of behavior and a change in the ways that neuronal activity summates. Furthermore, these bouts punctuate and characterize an exploitative behavioral state, whereby animals frequently enter periods of strong engagement in repeatedly performing a well-known action 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

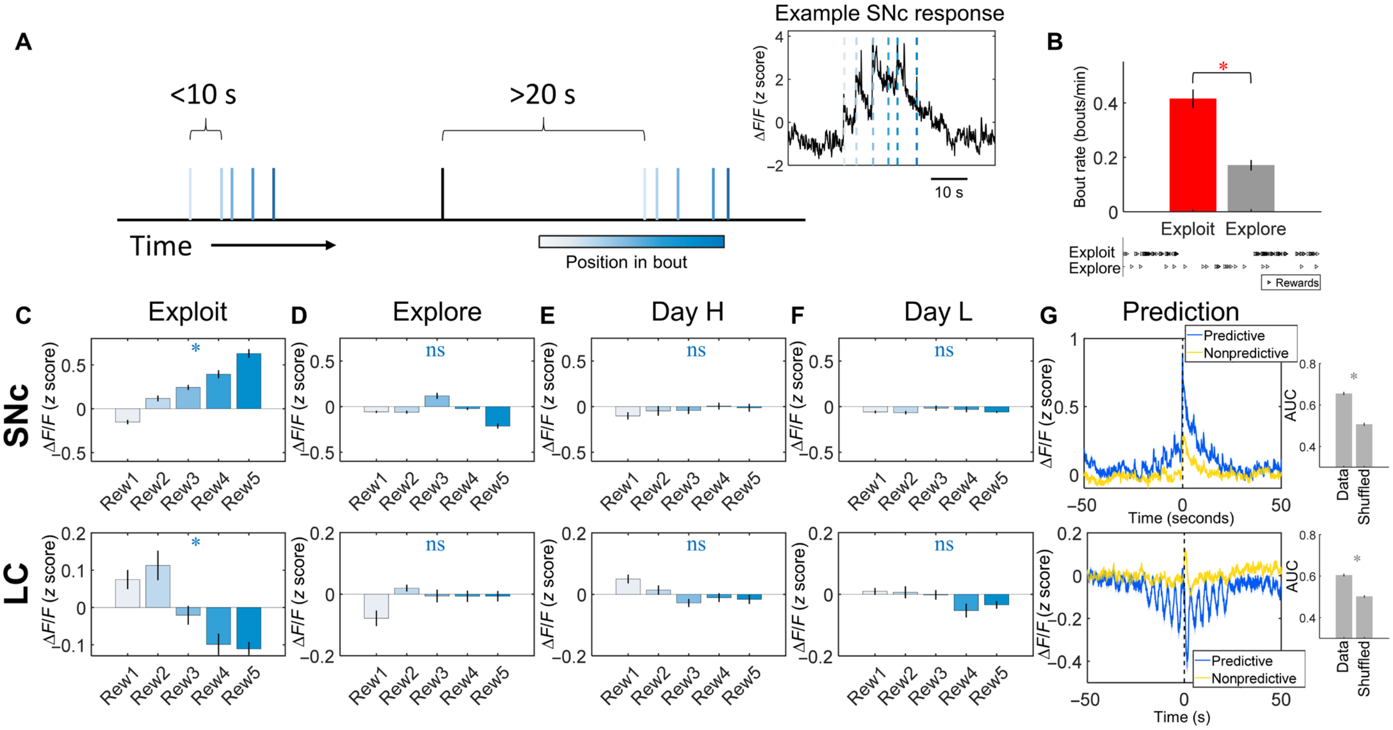

Fig. 2. Activity accumulates positively in SNc and negatively in LC during exploitative action-reward bouts. (**A**) Action-reward bouts are defined as clusters of action-reward pairs separated from each other by less than 10 s and separated from other action-reward pairs by more than 20 s. Right: Raw SNc fluorescence trace during an example action-reward bout, with individual action-reward pairs marked by dotted lines. (**B**) The mean rate of action-reward bout occurrence is increased in exploitative (red) relative to exploratory (gray) states. Bottom: Example raster showing occurrences of action-reward pairs clustering more into bouts during exploitation relative to exploration. (**C** to **F**) 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). (**G**) Left: Mean activity of neurons in SNc (top) and LC (bottom) whose activity is predictive (blue) or nonpredictive (yellow) of reward bouts using a wiener filter. Right: Mean across all cells of an area under the curve (AUC) assessment of wiener filter performance using activity from individual SNc (top) or LC (bottom) cells relative to prediction performance in cases in which the category labels were shuffled. Error bars denote SEM. *, significant at $P < 0.05$ or lower. ns, $P > 0.05$.

which activity accumulates over the course of action-reward bouts to produce sustained activity shifts. We therefore quantified the average duration and amplitude of all positive and negative neural activity transients (individual calcium events) during exploitative and exploratory epochs. We found an increase in the duration of positive response transients in SNc neurons during exploitative (6.89 ± 1.36 s) relative to exploratory (3.43 ± 0.46 s) behavioral states, resulting in extended response plateaus [$P = 0.002$, $t(4362) = 3.096$, and $N = 4364$ positive transients; Fig. 3A, left]. This increase in duration of response plateaus was not explained by changes in response magnitude [$P = 0.31$ and $t(4362) = 1.03$; Fig. 3A, right] and there was no change in the duration of negative response transients [$P = 0.24$, $t(680) = 1.17$, and $N = 682$ negative transients]. In addition, the lengthened positive response transients in SNc neurons during exploitative states were not observed during early learning (3.49 ± 0.20 s; fig. S6), on day H or day L (day H, 2.33 ± 0.42 s; day L, 2.09 ± 0.08 s; fig. S4), in the VTA (exploitation, 3.29 ± 0.21 s; exploration, 3.27 ± 0.07 s; fig. S7), or in the subset of SNc neurons whose activity was positively correlated with RL parameter estimates (fig. S12). In contrast, in LC neurons, we found no change in positive response transients across these behavioral states [$P = 0.09$, $t(5853) = 1.7$, and $N = 5855$ positive transients], but the duration of negative response transients was significantly longer in exploitative (4.67 ± 0.65 s) relative to exploratory (3.18 ± 0.30 s) states, resulting in extended response depressions [$P = 0.018$, $t(1504) = 2.37$, and $N = 1506$ negative transients; Fig. 3B, left]. This also was not explained by changes in response magnitude [$P = 0.552$ and $t(1504) = 0.63$; Fig. 3B, right]. Response transients from both regions were also fit individually with exponential functions, and with this metric, we again found an increase in the duration of positive transients in SNc [$P = 0.048$ and $t(588) = 1.98$] and negative transients in LC [$P = 0.023$ and $t(225) = 2.28$] during exploitative states, with no associated increase in the magnitude of transients (Fig. 3, C and D). To examine whether these activity changes in individual neurons were also reflected in changing network interactions, we analyzed the network correlation structure. Unexpectedly, we found activity in SNc cells to be more correlated across the network during action-reward bouts specifically during exploitative states, with smaller correlations during exploratory states or outside of action-reward bouts in exploitative states [interaction, $P = 0.0002$ and $F(1,1) = 14.1$; Fig. 3E]. LC cells were more correlated with each other during action-reward bouts relative to non-bouts in both exploitative and exploratory states [main effect, $P = 1.62 \times 10^{-7}$ and $F(1,1) = 27.61$; Fig. 3G]. To generate a more granular view of the consequences of these dynamics, we created cross-correlation histograms, where activity from all cells was time-locked to large fluorescence bursts in other simultaneously recorded cells. In the SNc during exploitative states, we found that cells in the network tended to increase activity together,

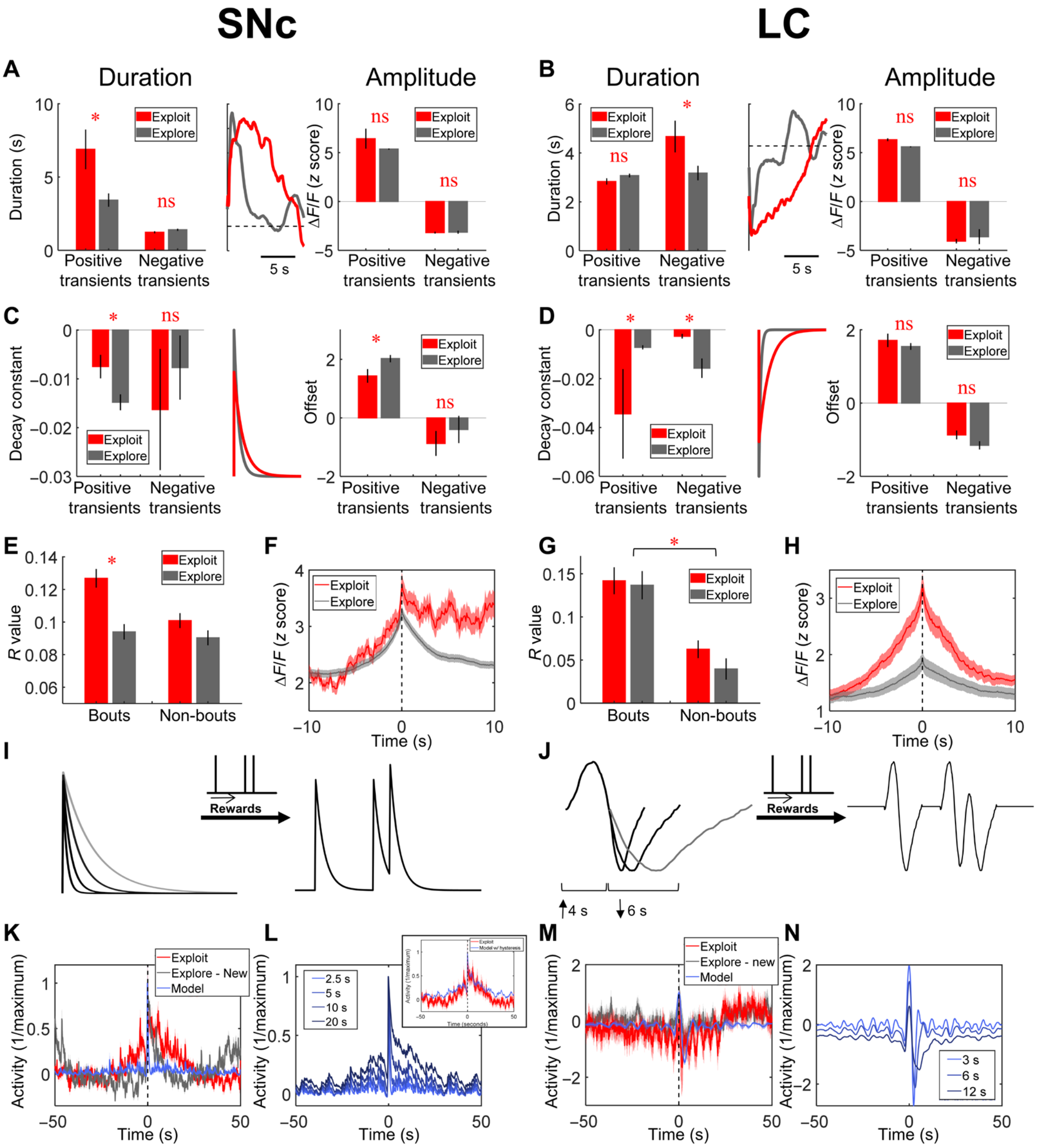

Fig. 3. Extended response plateaus in SNc and depressions in LC produce the sustained accumulation of activity. (**A** to **D**) Mean durations (left), amplitudes (right), and example traces (middle) of individual positive and negative response transients in SNc (A and C) and in LC (B and D) calculated directly from response transients (A and B) or from exponential fits to response transients (C and D) during exploitative (red) or exploratory (gray) states. (**E** and **G**) Mean peri-event pairwise activity correlations surrounding action-reward pairs that are in or out of action-reward bouts during exploitative (red) or exploratory (gray) states in SNc (E) or LC (G). (**F** and **H**) Cross-correlation histogram of mean neuronal activity in all cells time-locked to large fluorescence spikes in other simultaneously recorded cells during exploitative (red) and exploratory (gray) states in SNc (F) and LC (H). (**I** and **J**) Schematics showing impulse response functions (IRFs) used to produce reward convolution traces for SNc (I) and LC (J). (**K** to **N**) Reward convolution traces with typical IRFs in SNc (K) and LC (M), and with varying length IRFs in SNc (L) and LC (N), time-locked to reward achievement. Model responses in blue, original data for comparison in red (exploit) and gray (explore-new). Inset (L): Reward convolution traces with 10-s IRF and including hysteretic dynamics. Error bars denote SEM. *, significant at $P < 0.05$ or lower. ns, $P > 0.05$.

but many cells in the network then continued to fire afterward, exhibiting network-level hysteretic effects ($P < 0.001$; Fig. 3F). This asymmetry in the cross-correlation histogram suggests that the onsets of network responses in SNc are relatively synchronous, while the offsets are more asynchronous, with many cells continuing to respond with staggered offsets. In LC, correlated activity was generally increased during exploitation ($P < 0.01$), but the shape of this response was unchanged (Fig. 3H). The asymmetry seen in SNc during exploitative states was also not present in early learning (fig. S6), on day H or day L (fig. S4), in the VTA (fig. S7), or in the subset of SNc neurons whose activity was positively correlated with RL parameters (fig. S12). Dopaminergic and noradrenergic networks therefore both exhibit notable changes in response dynamics across exploitative and exploratory behavioral states, with dopaminergic populations also entering a regime of hysteretic network interactions.

To investigate whether these changing response dynamics could lead to the observed changes in sustained activity, we convolved the action-reward events in our behavioral data with an impulse response functions (IRFs) of various durations (Fig. 3, I and J). For SNc, this IRF was a simple exponential of varying length (Fig. 3I), while for LC, this IRF was the smoothed average population response to unexpected rewards (Fig. 3J). As we increased the duration of the SNc IRFs, we observed the emergence of sustained activity surrounding reward that matched that observed in SNc during exploitative states (Fig. 3, L and N). This was not observed with IRF durations more closely matched to the statistics of our neural data in baseline settings (Fig. 3, K and M). However, for IRFs matched to our data, the addition of hysteretic network dynamics to the model resulted in responses nearly identical to those seen in our data during exploitative states (Fig. 3L, inset, and fig. S13). The LC IRF, on the other hand, is biphasic, similar to classical responses seen with electrophysiology (*29*), with both a positive and negative phases that are asymmetric in baseline settings (Fig. 3J). If we alter the length of the negative phase of this IRF while holding the positive phase constant in our convolution model, then we again see the emergence of sustained baseline changes with longer IRFs (Fig. 3N). In both SNc and LC, the IRF durations modulate the degree to which responses to temporally proximal events, such as those seen in action-reward bouts, will summate. Together, these data suggest that the increased duration of positive transients (response plateaus) and network hysteresis in SNc, together with the increased duration of negative transients (response depressions) in LC, can result in differential summation of neuronal activity during exploitative and exploratory states, which, in turn, produced extended periods of enhanced dopaminergic activity and reduced noradrenergic activity during exploitative action-reward bouts.

Increasing dopaminergic or noradrenergic excitability differentially modulates motivation versus selection of ongoing actions

We next asked whether sustained changes in baseline activity levels could play a causal role in shifting between exploratory states and exploitative states, marked by action-reward bouts. Because these baseline shifts were due to changes in neural response dynamics that accumulate over the course of exploitative bouts, we used chemogenetic manipulations (hM3Dq) to enhance the excitability of genetically identified dopaminergic and noradrenergic populations (Fig. 4A) (*30*–*31*) rather than optogenetics to briefly drive excitability.

Animals expressing either hM3Dq or mCherry were imaged following intraperitoneal injections of either clozapine-*N*-oxide (CNO; the hM3Dq ligand) or vehicle (VEH) to determine the neuronal changes associated with chemogenetic activation. Following administration of CNO, we found a lengthening of the duration of positive response transients in SNc [$P = 5.79 \times 10^{-7}$ and $t(574) = 5.06$; Fig. 4, B and D] and LC [$P = 3.79 \times 10^{-5}$ and $t(30282) = 4.12$; Fig. 4, C and E] relative to VEH, as well as a shortening of the duration of negative response transients in LC [$P = 9.9 \times 10^{-4}$ and $t(3427) = 3.296$], similar to the neuronal changes we observed during natural behavior.

We therefore trained the animals on the task and noted the effects of CNO administration during exploitative states, exploratory states, and on day H. Exploitative action-reward bouts involve both the motivation to perform actions, similar to the hot-hand effect, and the selection of the same action to perform. Because sustained activity in SNc was related to response vigor and sustained activity in LC was not, we hypothesized that the motivation and selection aspects of exploitative states might be differentially mediated by dopaminergic and noradrenergic systems, respectively. During exploitative states, we found that enhancing LC excitability with CNO produced an increase in transition entropy [$P = 0.041$ and $t(7) = 2.49$] and a decrease in reward rate [$P = 0.043$ and $t(7) = 2.46$] relative to VEH, and this effect was not seen in control animals expressing mCherry (Fig. 4G, top row, and fig. S14). The same LC manipulation did not produce an effect on either entropy or reward rate during exploratory states, when animals still did not know the correct action sequence, or on day H, when animals performed many rewarded actions, showing that this was not a general disruption of behavior. Conversely, we found no change in entropy [$P = 0.159$ and $t(6) = 1.61$] or reward rate [$P = 0.618$ and $t(6) = 0.52$] when enhancing SNc excitability (Fig. 4F, top row, and fig. S14), suggesting that the diversity of actions to perform is mediated by the noradrenergic, but not dopaminergic, system. We therefore asked whether enhancing SNc excitability instead affected the motivation and structuring of action execution, with a particular focus on the clustering of behavior into action-reward bouts. We found little effect of enhancing SNc excitability on the proportion of rewards that occurred in action-reward bouts during exploitative states, when animals already perform a high proportion of actions in action-reward bouts [$P = 0.42$ and $t(4) = 0.89$; Fig. 4F, bottom row, and fig. S14]. Similarly, we found little difference in this measure when enhancing SNc excitability during exploratory states, before animals had learned which action sequence would lead to reward [$P = 0.21$ and $t(9) = 1.35$]. However, on day H, when animals could perform a wider range of actions to get the same high rate of reward, we found that enhancing SNc activity led to a significantly higher proportion of the total rewards occurring in action-reward bouts [$P = 0.046$ and $t(5) = 2.63$; effect was not present in controls expressing mCherry in SNc; Fig. 4F, bottom row, and fig. S14]. The mean number of action-reward pairs per action-reward bout also increased significantly following CNO [$P = 0.0041$ and $t(6) = 4.50$; fig. S14]. The emergence of this effect only on day H, when multiple actions can lead to reward, suggested that dopaminergic manipulations might be altering the motivation of action execution in a general manner rather than targeted toward specific actions alone. We therefore defined "engagement bouts" (see Materials and Methods) to quantify the clustering of all task-related actions in time regardless of whether they resulted in reward, in contrast to action-reward bouts that only capture the clustering of rewarded actions. We found an increase in the duration of engagement bouts following administration of CNO on day H [$P = 0.0125$ and $t(5) = 3.81$; Fig. 4F, bottom row, and fig. S14], suggesting that increasing SNc excitability enhanced the motivation to perform actions broadly in a non–action-specific manner (with high entropy). This restructuring of action-reward bouts and engagement bouts following CNO injections in animals expressing hM3Dq in SNc was not seen in animals expressing hM3Dq in LC (fig. S14), suggesting that the motivation for frequent action execution is mediated by the dopaminergic, but not noradrenergic, system. Enhanced noradrenergic activity therefore primarily affects levels of response entropy, driving transitions into exploratory behavioral states and increasing the diversity of actions to be executed, while enhanced dopaminergic activity primarily affects the motivation to execute actions, resulting in the restructuring of task performance into exploitative bouts of action.

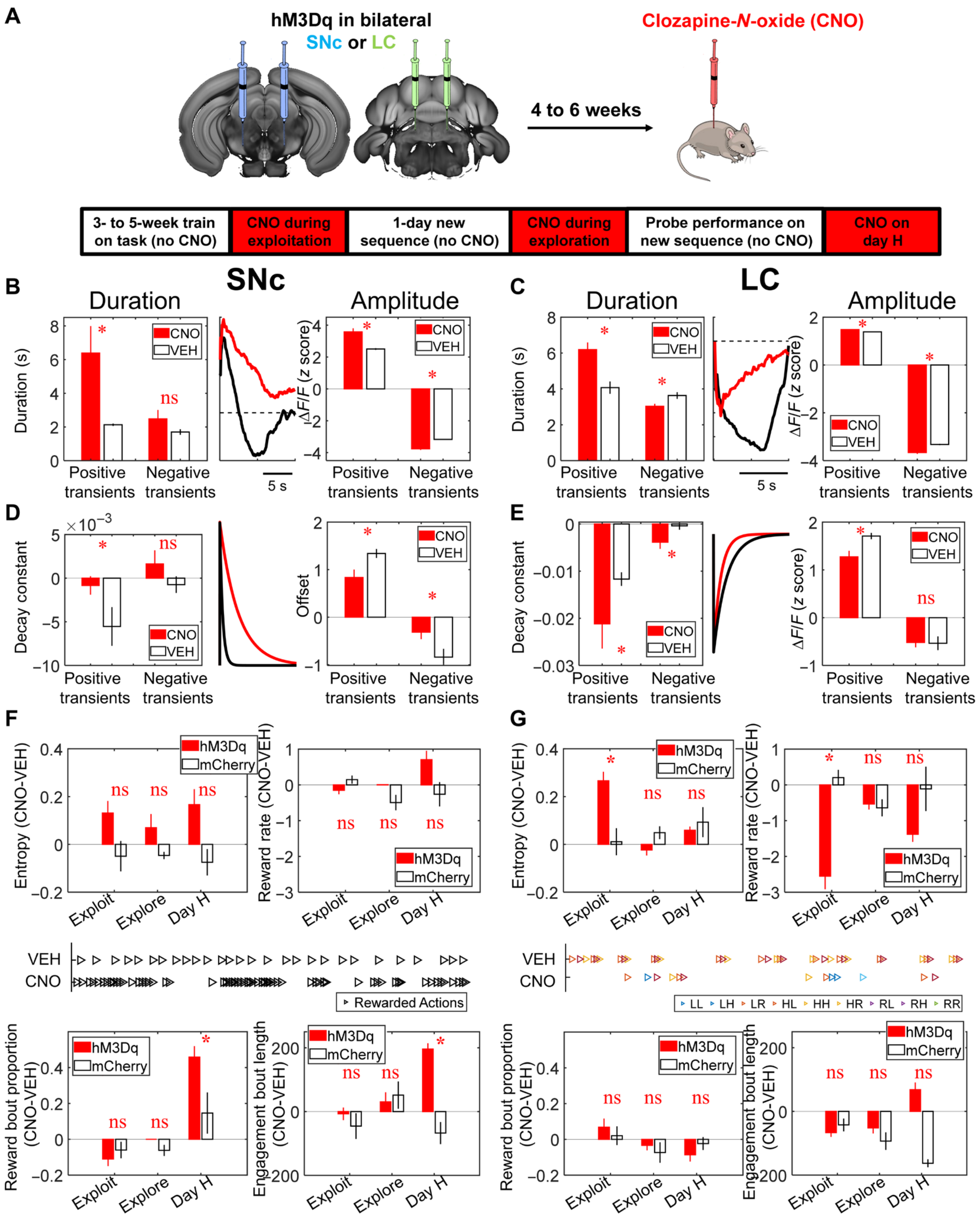

Fig. 4. Modulating excitability in dopaminergic and noradrenergic neurons differentially biases action motivation and selection. (**A**) Timeline for chemogenetic experiments. (**B** to **E**) Mean durations (left), amplitudes (right), and example traces (middle) of positive and negative transients in SNc (B and D) and in LC (C and E) calculated directly from response transients (B and C) or from exponential fits to response transients (D and E) following injection of CNO (red) or VEH (black). (**F**) Transition entropy (top left), reward rate (top right), proportion of action-reward pairs that occurred in bouts (bottom left), and duration of engagement bouts (bottom right) in animals expressing hM3Dq in SNc (red) and animals expressing mCherry in SNc (black), represented as the difference between values seen following CNO or following VEH (CNO-VEH). Middle: Example raster showing the timing and clustering of action-reward events following injections of VEH (top) or CNO (bottom). (**G**) Same measures as (F), but for animals expressing hM3Dq in LC (red) and animals expressing mCherry in LC (black). Middle: Example raster showing the selected transitions between nose pokes following injections of VEH (top) or CNO (bottom). Error bars denote SEM. *, significant at $P < 0.05$ or lower. ns, $P > 0.05$.

DISCUSSION

In conclusion, we developed a novel behavioral task in rodents to capture exploitative and exploratory states of action selection, and we found sustained changes in baseline activity in both dopaminergic and noradrenergic populations across these states. These sustained baseline changes in SNc were due to an increased duration of SNc response transients (plateaus), as well as hysteretic network dynamics, that together resulted in accumulations of activity and extended periods of enhanced dopaminergic activity during exploitative bouts, when animals repeatedly and vigorously performed well-learned actions to achieve a desired outcome. Conversely, the sustained inhibition in LC was caused by a lengthening of negative LC response transients (depressions) during exploitative states, resulting in progressively lower activity levels over the course of exploitative bouts. The response dynamics in both regions were strongly influenced not only by the occurrence of individual behavioral events or the macroscopic behavioral state (exploitative or exploratory) but also by the intermediate-scale action-reward bouts nested within these macroscopic states and composed of individual behavioral events. These altered response dynamics in SNc were not driven by subpopulations correlated with classical RL parameters and instead appeared to correspond to neuronal subpopulations coding for movement vigor, while sustained inhibition in LC was observed in subpopulations that do not code for movement vigor. Chemogenetic increases of excitability in dopaminergic and noradrenergic populations produced dissociable effects on the motivation and selection of actions, with the dopaminergic system primarily mediating the general motivation to perform actions and the noradrenergic system primarily affecting the diversity of specific actions to be performed. These results uncover two different aspects relating neural mechanisms to behavioral states. The first is that behavioral states are mediated by neural states that can last dozens of seconds and emerge as a result of subtle changes in the response properties of single neurons in a network that accumulate over intermediate time scales. The second is that dopaminergic and noradrenergic systems subserve dissociable aspects of exploitative behavioral states.

The response plateaus and depressions observed here with calcium imaging have a number of possible biological underpinnings. Most simply, these altered dynamics could reflect a change in the excitability or gain of individual neurons, which would be consistent with the observed effects following chemogenetic manipulations. This could, as one example, involve a change in the probability of neuronal up- or down-states across the population (*32*). Alternatively, these plateaus and depressions could reflect changing lateral interactions within the network (*33*), which is also consistent with the observed hysteretic network effects. More granular investigations with physiological methods are necessary to disentangle these possibilities.

The design of our task and experiments allowed us to investigate behavioral and neural states that occur on an intermediate time scale between the scale of synaptic signaling (milliseconds) and the scale of long-term potentiation (hours to days), a time scale that is much more similar to that experienced during ongoing behavior and perhaps more relevant to slower- and longer-acting neuromodulatory systems and the distributed dynamics on which they act. There are a number of ways that activity modulations on these intermediate time scales could differentially affect downstream circuits for action selection in the dorsal striatum, in the case of DA, and the anterior cingulate cortex, in the case of NE. Both systems are known to contain a range of receptor subtypes with distinct postsynaptic effects and behavioral correlates, and these downstream receptors have been hypothesized to respond preferentially to particular temporal dynamics of neuromodulator release (*34*–*39*). Further work is necessary to characterize the changes in target structures produced by neuromodulator systems during these intermediate-scale behavioral states.

Together, our results suggest that both dopaminergic and noradrenergic signaling modulate the likelihood of transitioning into an exploitative state: A state of inspired engagement in performing a well-known action that might be colloquially referred to as the hot-hand effect or being in the zone. Working in conjunction across multiple temporal scales, these systems structure our execution and selection of behaviors to either maintain a series of successes and capitalize on our learned skills or, conversely, to explore alternative actions and find novel, creative behavioral responses to an endlessly complex and nuanced environment.

MATERIALS AND METHODS

Animals

All experiments were performed in compliance with the regulations of the Institutional Animal Care and Use Committee at the Columbia University and the Ethics Board at the Champalimaud Centre for the Unknown. A total of 48 mice (13 female and 35 male) of roughly 3 months of age were used for the experiments. We saw no significant differences in basic behavior across the sexes (fig. S15) and therefore pooled data for all subsequent analyses. 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 DA 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 to 3%). Stereotactic coordinates relative to bregma were used to target the SNc (anteroposterior, −3.16 mm; mediolateral, ±1.4 mm; and dorsoventral, −4.2 mm) and stereotactic coordinates relative to lambda were used to target the LC (anteroposterior, −0.8 mm; mediolateral, ±0.8 mm; and dorsoventral, −3.2 mm). For imaging experiments, animals were injected unilaterally with 500 nl 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 nl of AAV5-hSyn-DIO-hM3D(Gq)-mCherry (Addgene plasmid no. 44361), while control animals were injected bilaterally with 500 nl of AAV5-hSyn-DIO-mCherry (Addgene plasmid no. 50459). All injections were performed using a Nanoject II Injector (Drummond Scientific, Broomall, PA, USA) at a rate of 4.6 nl every 5 s. Injection pipettes were left in place for 10 min after 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 GRIN lens (500 μm diameter, 8 mm length; 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). Two to three weeks were allowed for viral expression before attachment of microendoscope baseplates (Inscopix Inc.) to the dental acrylic at the correct focal plane for imaging.

Chemogenetics

For chemogenetic experiments, mice expressing either hM3Dq or mCherry were briefly anesthetized with 1 to 3% isoflurane and injected intraperitoneally with CNO or VEH (5 mg/kg) before behavioral sessions. Mice were given 15 min following injections to allow for the injection to take effect and for anesthetic effects to subside. Experiments were performed over the course of two sessions for each task condition (exploitation, exploration, and day H), with the order of CNO and VEH injections randomized across the cohort. Both hM3Dq and mCherry cohorts were tested following injections of CNO and VEH to control for nonspecific effects of CNO. Imaging the effects of hM3Dq activation was performed in the absence of task demands over the course of two sessions, randomized for CNO/VEH injection. A total of 25 mice were used for chemogenetics experiments, including 8 mice expressing hM3Dq in SNc, 9 mice expressing hM3Dq in LC, 4 mice expressing mCherry in SNc, and 4 mice expressing mCherry in LC.

Behavioral task

Animals were trained in custom-made operant boxes (5 inches by 6 inches) controlled by a Python-based framework (PyControl, https://pycontrol.readthedocs.io) that supplies all cues and rewards, as well as recording all behavioral time stamps. Behavior was also monitored with overhead cameras (Flea3, Point Grey Research, Richmond, Canada) recording at 30 frames per second. Operant boxes were placed inside sound attenuating chambers during training. Time stamps from the behavioral task were synchronized with calcium imaging data using TTL pulses sent from the behavioral chambers to the Inscopix data acquisition system.

Operant chambers contained three equidistant nose-poke ports surrounding a central reward port. Mice had to find a rewarded sequence of three pokes in a specific order with no intervening pokes. The task contains no trial structure and few cues, ensuring that mice actively explore the environment to find what is rewarding. When a correct sequence was performed, water rewards of 5 to 15 μl were supplied through the opening of a solenoid.

Mice were initially pretrained in a setting in which any possible three-poke sequence that includes all three nose-poke ports was rewarded. Following roughly 1 to 2 weeks of pretraining, mice were exposed to the full task, in which only one target sequence was rewarded. Once mice achieved proficiency on a particular target sequence (less than 10% continued improvement in reward rate across sessions), 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. Both day H and day L were performed after training on the main task and reversal sessions, and mice were given one full session to habituate to the new reward contingencies (separately for both day H and day L) before imaging data were collected for these conditions. Under all conditions, rewards could not be cached and had to be consumed before 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 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

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

where x_i represents either the full three-poke action sequences or the nose-poke transitions for calculating sequence entropy or transition entropy, respectively.

Corrections for finite sample sizes (*9*) 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 RL model was also applied to the behavioral data. Expected action values, Q, were updated on each trial according to

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

where c is the action chosen on trial t, δ_t is the RPE on trial t, and α is the learning rate of the model. Expected action values were related to choices by the following equation

$$p_t(c) = \frac{e^{\beta Q_t(c)}}{\sum_{b=1}^{n} e^{\beta Q_t(b)}}$$

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

$$V_t(s_t) = \sum_{a=1}^{n} Q_t(c)\,p_t(c)$$

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

Behavioral data were also separated into action-reward bouts and engagement bouts. "Action-reward bouts" were defined as groups of multiple action-reward events (where animals performed the target action sequence and were rewarded for it) that were separated from each other by less than 10 s and separated from other action-reward events by more than 20 s. "Engagement bouts," on the other hand, were defined as groups of multiple individual actions (single-task pokes rather than target three-poke sequences) that were separated from each other by less than 2 s and separated from other actions by more than 5 s, irrespective of whether or not the actions resulted in reward.

Calcium imaging data were first preprocessed using Mosaic (Inscopix Inc.) to apply 4× spatial downsampling and motion correction. CNMF-E (*20–21*) was then applied for demixing and further preprocessing of the data. The footprints and activity profiles of all putative neurons were inspected manually before inclusion. Because of our interest in slowly varying baseline activity fluctuations, the CNMF-E output C_raw, corresponding to a scaled version of the conventional $\Delta F/F$, was used for all analyses rather than the output C. Output traces were then z scored before all analyses. A total of 121 cells in SNc (7 animals, mean 23 cells per animal), 61 cells in LC (10 animals, mean 6.1 cells per animal), and 9 cells in VTA (2 animals, mean 4.5 cells per animal) were included in the primary analyses. When multiple comparisons were performed, corrections were made using the false discovery rate.

A Wiener filter was trained to discriminate action-reward pairs occurring within bouts from action-reward pairs occurring outside of bouts. For each action-reward pair, five lags were used occurring every 500 ms starting 2 s before the action-reward pair. 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 curve. Chance performance was assessed by testing predictive performance when the behavioral category labels were shuffled. "Predictive cells" were defined as cells with AUC > 0.6, while "nonpredictive cells" were defined as cells with AUC < 0.5.

To separate vigor-correlated and non–vigor-correlated neuronal populations, a median split was performed on the data based on the correlation coefficient with the negative of the inter-poke interval. The neurons most correlated with fast, vigorous poking were considered vigor-correlated neurons and all others were considered non–vigor-correlated.

For convolution models, pure exponentials of varying lengths were used to model the SNc IRF. For the LC IRF, the average response from all LC cells to unexpected rewards was smoothed by a 1-s moving average. To model hysteretic network dynamics, multiple convolution traces were created for each animal with the addition of a second neural response following the initial reward-locked response by a random fraction of a maximum of 5 s.

To quantify response transient durations, positive and negative threshold crossings (3SD) were located. A 5-s window before threshold crossing was defined as baseline activity, and a 5-s 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 time points by which the second window had to be advanced to equal the baseline window was defined as the response transient duration. Individual transients were also fit with exponential functions to determine decay and offset parameters.

SUPPLEMENTARY MATERIALS

Supplementary material for this article is available at http://advances.sciencemag.org/cgi/content/full/7/30/eabh2059/DC1

View/request a protocol for this paper from *Bio-protocol*.

REFERENCES AND NOTES

1. F. Cinotti, V. Fresno, N. Aklil, E. Coutureau, B. Girard, A. R. Marchand, M. Khamassi, Dopamine blockade impairs the exploration-exploitation trade-off in rats. *Sci. Rep.* **9**, 6770 (2019).
2. K. Chakroun, D. Mathar, A. Wiehler, F. Ganzer, J. Peters, Dopaminergic modulation of the exploration/exploitation trade-off in human decision-making. *eLife* **9**, e51260 (2020).
3. R. M. Costa, Plastic corticostriatal circuits for action learning: What's dopamine got to do with it? *Ann. N. Y. Acad. Sci.* **1104**, 172–191 (2007).
4. G. Aston-Jones, J. D. Cohen, An integrative theory of locus coeruleus-norepinephrine function: Adaptive gain and optimal performance. *Annu. Rev. Neurosci.* **28**, 403–450 (2005).
5. M. Usher, J. D. Cohen, D. Servan-Schreiber, J. Rajkowski, G. Aston-Jones, The role of locus coeruleus in the regulation of cognitive performance. *Science* **283**, 549–554 (1999).
6. A. Izquierdo, L. M. Wiedholz, R. A. Millstein, R. J. Yang, T. J. Bussey, L. M. Saksida, A. Holmes, Genetic and dopaminergic modulation of reversal learning in a touchscreen-based operant procedure for mice. *Behav. Brain Res.* **171**, 181–188 (2006).
7. M. Klanker, M. Feenstra, D. Denys, Dopaminergic control of cognitive flexibility in humans and animals. *Front. Neurosci.* **7**, 201 (2013).
8. M. D. Humphries, M. Khamassi, K. Gurney, Dopaminergic control of the exploration-exploitation trade-off via the basal ganglia. *Front. Neurosci.* **6**, 9 (2012).
9. D. G. R. Tervo, M. Proskurin, M. Manakov, M. Kabra, A. Vollmer, K. Branson, A. Y. Karpova, Behavioral variability through stochastic choice and its gating by anterior cingulate cortex. *Cell* **159**, 21–32 (2014).
10. M. P. Karlsson, D. G. R. Tervo, A. Y. Karpova, Network resets in medial prefrontal cortex mark the onset of behavioral uncertainty. *Science* **338**, 135–139 (2012).
11. J. D. Cohen, S. M. McClure, A. J. Yu, Should I stay or 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 (2007).
12. P. Dayan, N. D. Daw, Decision theory, reinforcement learning, and the brain. *Cogn. Affect. Behav. Neurosci.* **8**, 429–453 (2008).
13. N. D. Daw, Y. Niv, P. Dayan, Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioral control. *Nat. Neurosci.* **8**, 1704–1711 (2005).
14. R. B. Ebitz, E. Albarran, T. Moore, Exploration disrupts choice-predictive signals and alters dynamics in prefrontal cortex. *Neuron* **97**, 450–461.e9 (2018).
15. M. Speekenbrink, E. Konstantinidis, Uncertainty and exploration in a restless bandit problem. *Top. Cogn. Sci.* **7**, 351–367 (2015).
16. B. Y. Hayden, J. M. Pearson, M. L. Platt, Neuronal basis of sequential foraging decisions in a patchy environment. *Nat. Neurosci.* **14**, 933–939 (2011).
17. T. Neiman, Y. Loewenstein, Reinforcement learning in professional basketball players. *Nat. Commun.* **2**, 569 (2011).
18. T. C. Blanchard, A. Wilke, B. Y. Hayden, Hot-hand bias in rhesus monkeys. *J. Exp. Psychol. Anim. Learn. Cogn.* **40**, 280–286 (2014).
19. A. K. Dhawale, M. A. Smith, B. P. Olveczky, The role of variability in motor learning. *Annu. Rev. Neurosci.* **40**, 479–498 (2017).
20. P. Zhou, S. L. Resendez, J. Rodriguez-Romaguera, J. C. Jimenez, S. Q. Neufeld, A. Giovanucci, J. Friedrich, E. A. Pnevmatikakis, G. D. Stuber, R. Hen, M. A. Kheirbek, B. L. Sabatini, R. E. Kass, L. Paninski, Efficient and accurate extraction of in vivo calcium signals from microendoscopic video data. *Elife* **7**, e28728 (2018).
21. E. A. Pnevmatikakis, D. Soudry, Y. Gao, T. A. Machado, J. Merel, D. Pfau, T. Reardon, Y. Mu, C. Lacefield, W. Yang, M. Ahrens, R. Bruno, T. M. Jessell, D. S. Peterka, R. Yuste, L. Paninski, Simultaneous denoising, deconvolution, and demixing of calcium imaging data. *Neuron* **89**, 285–299 (2016).
22. A. Klaus, G. J. Martins, V. B. Paixao, P. Zhou, L. Paninski, R. M. Costa, The spatiotemporal organization of the striatum encodes action space. *Neuron* **96**, 949 (2017).
23. A. A. Hamid, J. R. Pettibone, O. S. Mabrouk, V. L. Hetrick, R. Schmidt, C. M. Vander Weele, R. T. Kennedy, B. J. Aragona, J. D. Berke, Mesolimbic dopamine signals the value of work. *Nat. Neurosci.* **19**, 117–126 (2016).
24. W. Schultz, P. Dayan, P. R. Montague, A neural substrate of prediction and reward. *Science* **275**, 1593–1599 (1997).
25. J. A. da Silva, F. Tecuapetla, V. Paixao, R. M. Costa, Dopamine neuron activity before action initiation gates and invigorates future movements. *Nature* **554**, 244–248 (2018).
26. M. W. Howe, P. L. Tierney, S. G. Sandberg, P. E. M. Phillips, A. M. Graybiel, Prolonged dopamine signalling in striatum signals proximity and value of distant rewards. *Nature* **500**, 575–579 (2013).
27. A. Guru, C. Seo, R. J. Post, D. S. Kullakanda, J. A. Schaffer, M. R. Warden, Ramping activity in midbrain dopamine neurons signifies the use of a cognitive map. *bioRxiv* 10.1101/2020.05.21.108886 , (2020).
28. J. G. Mikhael, H. R. Kim, N. Uchida, S. J. Gershman, Ramping and state uncertainty in the dopamine signal. *bioRxiv* 10.1101/805366 , (2019).
29. G. Aston-Jones, J. Rajkowski, J. D. Cohen, Role of locus coeruleus in attention and behavioral flexibility. *Biol. Psychiatry* **46**, 1309–1320 (1999).
30. B. N. Armbruster, X. Li, M. H. Pausch, S. Herlitze, B. L. Roth, 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**, 5163–5168 (2007).
31. D. J. Urban, B. L. Roth, DREADDs (designer receptors exclusively activated by designer drugs): Chemogenetic tools with therapeutic utility. *Annu. Rev. Pharmacol. Toxicol.* **55**, 339–417 (2015).
32. D. Jercog, A. Roxin, P. Bartho, A. Luczak, A. Compte, J. de la Rocha, UP-DOWN cortical dynamics reflect state transitions in a bistable network. *Elife* **6**, e22425 (2017).
33. A. Sahasranamam, I. Vlachos, A. Aertsen, A. Kumar, Dynamical state of the network determines the efficacy of single neuron properties in shaping the network activity. *Sci. Rep.* **6**, 26029 (2016).
34. A. V. Kravitz, L. D. Tye, A. C. Kreitzer, Distinct roles for direct and indirect pathway striatal neurons in reinforcement. *Nat. Neurosci.* **15**, 816–818 (2012).
35. F. Tecuapetla, X. Jin, S. Q. Lima, R. M. Costa, Complementary contributions of striatal projection pathways to action initiation and execution. *Cell* **166**, 703–715 (2016).
36. J. K. Dreyer, K. F. Herrik, R. W. Berg, J. D. Hounsgaard, Influence of phasic and tonic dopamine release on receptor activation. *J. Neurosci.* **30**, 14273–14283 (2010).
37. C. W. Berridge, R. C. Spencer, Differential cognitive actions of norepinephrine a2 and a1 receptor signaling in the prefrontal cortex. *Brain Res.* **1641**, 189–196 (2016).
38. S. Lohani, A. K. Martig, K. Deisseroth, I. B. Witten, B. Moghaddam, Dopamine modulation of prefrontal cortex activity is manifold and operates at multiple temporal and spatial scales. *Cell Rep.* **27**, 99–114.e6 (2019).

39. A. F. Arnsten, Through the looking glass: Differential noradenergic modulation of prefrontal cortical function. *Neural Plast.* **7**, 133–146 (2000).

Acknowledgments: We thank V. Athalye and J. Murray for the insightful discussion. **Funding:** This material is based on work supported by the NIH 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. **Author contributions:** A.C.K. and R.M.C. conceived and designed experiments and analyses. A.C.K. performed data collection and analysis. A.C.K. and R.M.C. prepared the manuscript. **Competing interests:** The authors declare that they have no competing interests. **Data and materials availability:** All data needed to evaluate the paper are present in the paper and/or the Supplementary Materials. Additional data related to this paper may be requested from the authors.

Submitted 22 February 2021
Accepted 7 June 2021
Published 23 July 2021
10.1126/sciadv.abh2059

Citation: A. C. Koralek, R. M. Costa, Dichotomous dopaminergic and noradrenergic neural states mediate distinct aspects of exploitative behavioral states. *Sci. Adv.* **7**, eabh2059 (2021).