Cell

Behavioral Variability through Stochastic Choice and Its Gating by Anterior Cingulate Cortex

Dougal G.R. Tervo,[1] Mikhail Proskurin,[1] Maxim Manakov,[1] Mayank Kabra,[1] Alison Vollmer,[1] Kristin Branson,[1] and Alla Y. Karpova[1,*]
[1]Howard Hughes Medical Institute, Janelia Research Campus, 19700 Helix Drive, Ashburn, VA 20147, USA
*Correspondence: alla@janelia.hhmi.org
http://dx.doi.org/10.1016/j.cell.2014.08.037

SUMMARY

Behavioral choices that ignore prior experience promote exploration and unpredictability but are seemingly at odds with the brain's tendency to use experience to optimize behavioral choice. Indeed, when faced with virtual competitors, primates resort to strategic counterprediction rather than to stochastic choice. Here, we show that rats also use history- and model-based strategies when faced with similar competitors but can switch to a "stochastic" mode when challenged with a competitor that they cannot defeat by counterprediction. In this mode, outcomes associated with an animal's actions are ignored, and normal engagement of anterior cingulate cortex (ACC) is suppressed. Using circuit perturbations in transgenic rats, we demonstrate that switching between strategic and stochastic behavioral modes is controlled by locus coeruleus input into ACC. Our findings suggest that, under conditions of uncertainty about environmental rules, changes in noradrenergic input alter ACC output and prevent erroneous beliefs from guiding decisions, thus enabling behavioral variation.

INTRODUCTION

When an animal repeatedly encounters the same situation, its behavioral choices often vary, even when the optimal choice should be clear from past experience. The fact that an identical state of the environment can elicit different behavioral responses is often interpreted as evidence that variability in behavior is the unintended by-product of errors in decision making (Beck et al., 2012; Faisal et al., 2008; Gold and Shadlen, 2007; Padoa-Schioppa, 2013). But when uncertainty about environmental conditions favors exploration, animals may benefit from intentionally imposing variability on behavioral choices (Cohen et al., 2007; Ölveczky et al., 2005; Page and Neuringer, 1985; Sutton and Barto, 1998). One potential strategy for generating variability involves dispensing with prior beliefs, for example, by opting for a purely stochastic mode of action selection. Implementing such a strategy poses, however, a significant challenge in that the brain has to ignore widespread signals that normally encode past experience and rules of behavior derived from it (Buschman et al., 2012; Vickery et al., 2011). How then might the brain's ability to ignore past experience be exposed? A strategy of imposed variability might be favored in situations in which prediction of one's actions by a competitor or predator has adverse consequences (Nash, 1950; Maynard Smith and Harper, 1988)—a concept that can be captured experimentally by confronting the subject with an electronic competitor that strives to predict future choice on the basis of past behavior, rewarding only when prediction is eluded (Abe and Lee, 2011; Barraclough et al., 2004; Dorris and Glimcher, 2004; Lee et al., 2004, 2005). Nevertheless, construction of an internal mental model that effectively discerns the workings of a competitor could generate a successful counterpredictive strategy—a mental simulation of which prediction the competitor is likely to make—without the need for stochasticity. Indeed, studies in primates support the idea that competition triggers model-based counterprediction rather than stochastic choice (Abe and Lee, 2011; Zhu et al., 2012).

As such, it remains unclear whether the brain possesses the ability to implement stochastic action choice, rather than to rely exclusively on experience-derived models of the environment. Studies in which task rules change suddenly have provided clues that internal models can be overridden. In such settings, animals respond to rule changes by abruptly initiating exploratory behavior (Daw et al., 2006; Karlsson et al., 2012; Nassar et al., 2010; O'Reilly et al., 2013), implying the existence of a mechanism that can release behavioral control from the influence of an internal model that has been deemed inadequate. The construction of internal models, notably also in competitive settings, is thought to recruit neural activity in the anterior cingulate cortex (ACC), among other brain regions (Hayden et al., 2011; Holroyd and Yeung, 2012; Karlsson et al., 2012; Matsumoto et al., 2003; Ribas-Fernandes et al., 2011; Yoshida and Ishii, 2006; Zhu et al., 2012). This fits with the observation of widespread and coordinated changes in the activity of the neuronal population in the ACC concurrent with the decision to abandon an inadequate model and initiate exploration (Karlsson et al., 2012). It is thought that activation of the noradrenergic system signals the decision to abandon an inadequate model, driven presumably by unexpected mismatches between the internal model's predictions and environmental feedback (Jepma

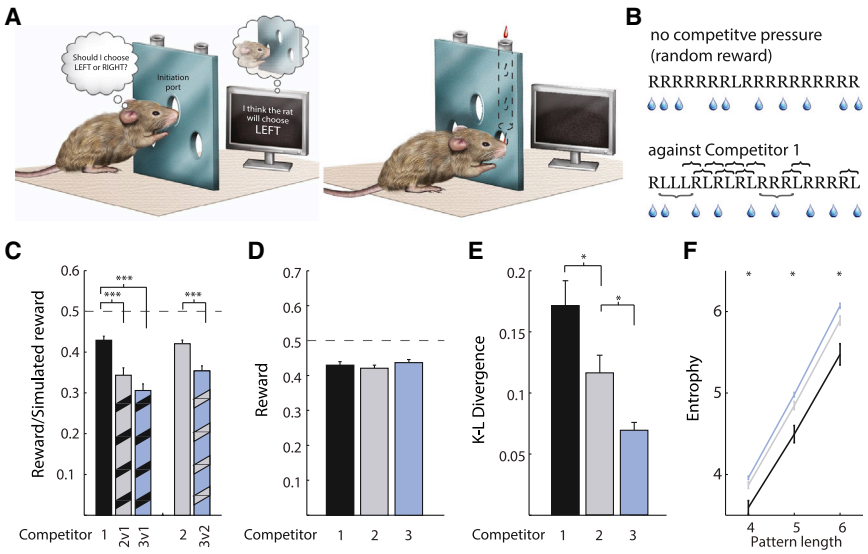

Figure 1. Rats' Behavior in a Virtual Competitive Environment

(A) Concept of the behavioral task. After initiating a trial at the central port, an animal is eligible to receive a reward only if his choice of the reward port differs from that predicted by the computer competitor. The three holes in the wall represent the initiation port (center) and both choice ports (left and right).

(B) Two example sequences of left (L) and right (R) choices from rats who face no competitive pressure when the computer chooses the reward port randomly and competitor 1, top and bottom, respectively, together with outcome (reward, indicated by the droplet, or no reward). Horizontal curly brackets indicate common patterns.

(C) Real reward (solid bars) in actual play and fictive reward (striped bars) in simulated play.

(D) First-session mean reward rates against the various competitors. The dashed line indicates the reward rate that rats would receive if their choices were generated by an unbiased stochastic process.

(E) Mean Kullbach-Leibler (K-L) divergence of the rats' behavior against the various competitors from optimal (see Experimental Procedures).

(F) Mean Shannon entropy of choice patterns against the various competitors. The significance of the difference between competitor 2 (gray line) and competitor 3 (blue line) is indicated by an asterisk. (C–F) n = 12 against competitor 1, n = 13 against competitor 2, and n = 12 against competitor 3. *p < 0.05, ***p < 0.001, Wilcoxon rank sum. Error bars represent the SEM.

See also Figure S1.

and Nieuwenhuis, 2011; Nassar et al., 2012; Payzan-LeNestour et al., 2013; Yu and Dayan, 2005), raising the possibility that modulation of ACC's model-encoding circuits by the input from the noradrenergic system underlies the release of behavior from the control of an internal model.

Here, we probe whether a stochastic strategy is adopted when behavioral choice is released from the influence of an internal model in complex settings, and whether the noradrenergic system plays a role in this behavioral switch, by taking advantage of regimes of sustained behavioral variability induced by competitive settings. Using a virtual competitive task, we test whether rats are still capable of generating variable behavioral choices when faced with a competitor that is sophisticated enough to thwart the animal's modeling attempts. We find that when faced with a competitor that they cannot defeat by counterprediction, animals switch to a distinct mode of action selection consistent with stochastic choice. In this mode, characterized by highly variable choice sequences, behavior becomes dramatically less dependent on the history of outcomes associated with different actions and becomes independent from the ACC. Moreover, selective enhancement or suppression of locus coeruleus input into the ACC, respectively, abolishes or restores model-based control of behavior and, with it, sensitivity to environmental feedback. Our findings argue that neural mechanisms for purposeful behavioral variability do exist and strongly suggest that noradrenergic action in ACC controls the extent to which behavioral choices are guided by the internal model or stochastic selection.

RESULTS

To explore whether stochastic choice can be exposed in a competitive setting, we trained rats on a task that required them to select one of two reward ports while being monitored by a computer-simulated (virtual) competitor (Figure 1A). The computer was programmed to search the history of animal performance for behavioral patterns that could be used to predict its upcoming choice. In this scheme, the animal is eligible to receive a reward at the chosen port only if its choice differs from that predicted by the computer. We first determined whether rats, like primates, use a counterpredictive strategy when they encounter a weak competitor. Specifically, we exposed rats to a virtual competitor that counteracts an animal's bias for selecting one of the two ports following a particular immediate history of choices and reward but only when this bias exceeds a preset threshold (see the Experimental Procedures). Against this competitor (competitor 1, similar to the one used in previous primate studies [Barraclough et al., 2004; Lee et al., 2004]), rats were able to make their choices sufficiently variable to sustain a relatively high average reward rate (41.6% ± 1.4%; Figure 1D), which rose further during subsequent sessions, sometimes surpassing (see below) the 50% expected for an unbiased stochastic strategy. Nevertheless, each rat's behavior still contained clearly detectable structure (see Figure 1B for a representative example; Figure S1A available online). This might simply reflect the animals' natural preference for simple patterns, something that is not completely suppressed by the competitive pressure applied by competitor 1, or indicate the use of counterprediction.

To distinguish between these two possibilities, we relied on the fact that an effective counterprediction strategy can actually lead to a reward rate that exceeds that for stochastic choice (50%). In fact, for some animals playing against competitor 1, we observed reward rates significantly higher than would be expected by chance deviations from 50%, reaching as high as 60%

($p < 0.001$ after Bonferroni correction for multiple comparisons; bootstrap from a binomial process against competitor 1). Thus, these animals were eventually able to model aspects of the underlying prediction algorithm and used that knowledge for counterprediction. A more sophisticated opponent is therefore necessary to thwart attempts at model construction.

More Sophisticated Electronic Competitors

To render feedback- and model-based strategies ineffective, we designed two more challenging competitors (2 and 3). Competitor 2 uses the same prediction algorithm as competitor 1, except it removes the requirement for the bias in favor of one side over the other to reach a predetermined threshold before competitive pressure is applied (see the Experimental Procedures). Competitor 3 uses a sophisticated machine-learning method, known as boosting (Friedman et al., 2000, see the Experimental Procedures), an algorithm of much greater complexity that learns to generate a strong prediction on the basis of a set of weak trends in the data.

We used simulated play to test whether the algorithms of competitors 2 and 3 provided better prediction than did competitor 1 by calculating the reward accrued from using animals' real behavioral performance against one competitor as simulated choices against another (see the Experimental Procedures). A stronger competitor would detect some of the patterns that a weaker competitor missed, leading to a correct prediction of the choices made by the animal—and with that, to a withholding of the reward—on more trials. The simulated reward against competitors 2 and 3 (34.4% and 30.6%) indeed fell short of the actual reward against competitor 1 (41.6%; Figure 1C; Wilcoxon rank sum, $p < 0.001$ for competitors 2 versus 1 and 3 versus 1, $n = 12$ animal histories; also see the Experimental Procedures). Furthermore, competitor 3 beat competitor 2 in simulated play (Figure 1C; 35.42% ± 1.24% simulated reward against competitor 3 versus 42.95% ± 1.03% actual reward against competitor 2, $p < 0.001$ Wilcoxon rank sum, $n = 13$ animal histories). These findings imply that competitors 2 and 3 can detect patterns left unpunished by competitor 1 and will, therefore, exert stronger competitive pressure requiring progressively more sophisticated counterpredictive strategies in order to sustain a high rate of reward. A stochastic strategy for producing variable behavioral choices would, of course, be effective against any of the competitors (1, 2, and 3).

Successful Performance against Stronger Competitors

Having established that competitors 2 and 3 exert stronger competitive pressure, we next investigated whether rats are still able to find successful strategies when actually playing against them. Animals accrued comparable first-session reward rates across all competitors (Figure 1D; not significant for competitor 1 versus competitor 2 or 3, Wilcoxon rank sum), suggesting that they produced fewer detectable patterns as the competitive pressure increased. We also tested the presence of patterns directly by comparing the conditional probabilities of choosing either the left or right reward port given a particular history pattern and found that they became more balanced for all patterns—a prerequisite for being optimally unpredictable (Figure 1E; $p < 0.05$ for Kullback-Leibler [K-L] divergence from optimal for competitors 2 versus 1, $n = 12$ animals; $p < 0.05$ for competitors 3 versus 2, $n = 13$ animals, Wilcoxon rank sum; see the Experimental Procedures). Furthermore, the animals' patterns of sequential choices were more uniformly distributed across the space of all possible patterns for greater competitive strengths, even for longer patterns (Figure 1F; Wilcoxon rank sum, $p < 0.05$ for entropy of choice sequences for competitors 2 versus 1 and competitors 3 versus 2 at pattern lengths 4, 5, and 6; see the Experimental Procedures). The previously used virtual competitor (competitor 1) thus did not reveal the brain's full capacity for generating behavioral variability.

Is the increase in behavioral-choice variability against competitors 2 and 3 indicative of an increase in the sophistication of counterprediction, or have such strategies been abandoned and replaced by the active generation of behavioral variability in a manner that is feedback- and model-independent? Multiple linear regression of the rats' choices on the choices and outcomes of the preceding three trials suggested that the strategy against competitor 2, but likely not against competitor 3, was still dependent on feedback from the environment (Figure S1B). We, therefore, set out to test directly whether the sophistication of competitor 3 was enough to push the animals into a feedback- and model-independent behavioral mode.

In designing an experimental approach to test whether the animals' choices depend on environmental feedback and model-based counterprediction, we reasoned that any behavioral mode that relies on feedback for generating variable behavior should also make the animal sensitive to changes in the statistics of the reward associated with different choice patterns, enabling it to detect and exploit novel opportunities in the environment. Furthermore, a behavioral strategy that involves the mental simulation of the competitor's prediction algorithm would be dependent on computations that likely take place in the ACC (Zhu et al., 2012). An effectively stochastic mode would, on the other hand, make behavior insensitive to environmental feedback and independent of model-related computations in the ACC.

Adoption of a Strategy that Ignores Environmental Feedback

We first assessed whether animals come to ignore the correlation between behavioral patterns and environmental feedback when faced with stronger competitive pressure. To test this, we surreptitiously switched animals that had been playing against competitors 2 or 3 to a specifically designed task that requires an animal to discover that a particular (covert) pattern of choices is always rewarded but do so in the presence of a parallel reward stream that weakly encourages unpredictable rather than structured behavior (Figure 2A; see the Experimental Procedures). We chose one of two three-step patterns (left-left-right, "LLR" or right-right-left, "RRL," neither of which occur during simple alternation) and set the reward rates such that adhering to that pattern would be significantly more beneficial than maintaining highly variable behavior (Experimental Procedures). In this "covert pattern" task, rats are rewarded in a fraction of trials even if their choices do not conform to the covert pattern (~16% of trials if the animal's behavior is fully random). Nevertheless, animals that remain sensitive to the reward statistics associated

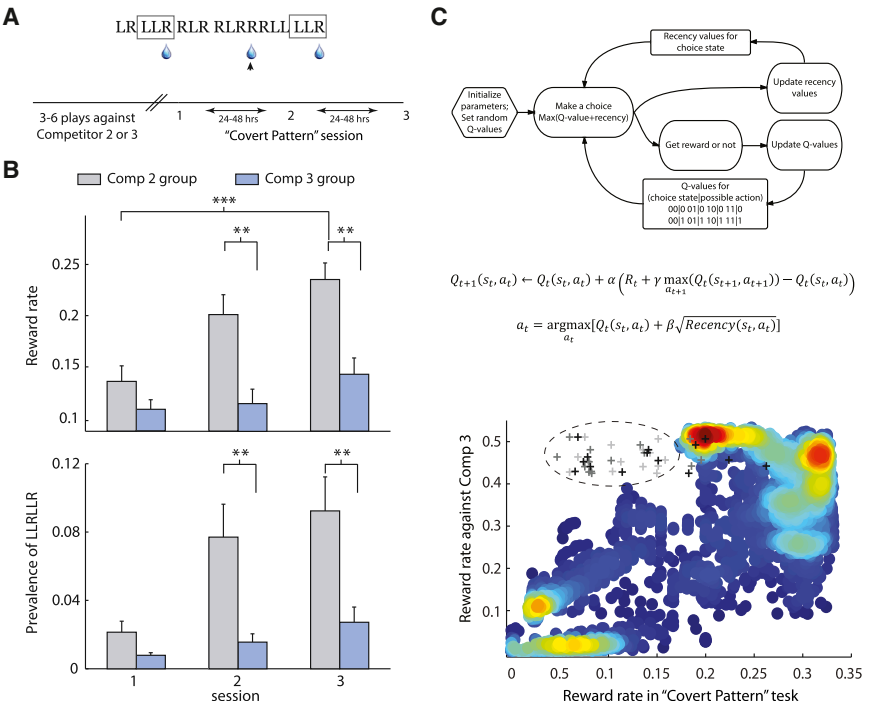

Figure 2. Competitor 3 Induces Insensitivity to Feedback that Is Revealed by "Covert Pattern" Task

(A) Boxed LLR indicates the sequence that receives reward (droplet). The arrow points to a reward received due to the background competitor.

(B) Reward rates (top) and prevalences of LLRLLR (two covert sequences in a row, bottom) in the three probe sessions following the switch from competitors 2 (gray, comp 2) and 3 (blue, comp 3) to the covert sequence task. n = 8 animals for competitor 2 group; n = 16 animals for competitor 3 group.

(C) Performance of Q-learners. Top: Q-learning reinforcement learning algorithm (schematic). Bottom: density plot across all parameter values of Q-learners' performance against Competitor 3 (comp 3) versus performance in the "covert pattern" task. Hotter colors represent higher density. Crosses indicate actual rat performance in competitor 3 group for comparison. Light gray, first LLR session after switch from the competitor 3 setting. Dark gray, second LLR session. Black, third LLR session. Dashed ellipse highlights the subspace that contains most of the rat data points but is devoid of coverage by Q-learner performance. **p < 0.01; ***p < 0.001, Wilcoxon rank sum. Error bars represent the SEM.

with particular sequences of actions should eventually be able to infer that biasing their behavioral choices to conform to the covert pattern provides the greatest reward (with 33% being the maximum achievable for a perfect concatenation of LLR's or RRL's). In contrast, animals that switch to a feedback- and model-independent strategy would be less likely to discover this opportunity to increase the reward rate.

Rats that had beforehand been playing against competitor 2 substantially improved in their ability to procure reward in this "covert pattern" task over the course of the first three sessions. Most animals in this group reached reward rates above 20% by the third session (Figure 2B, top; Wilcoxon rank sum, $p < 0.001$ for session 1 versus 3), which was similar to the "expert" level of performance on this task observed after over 15 sessions of training (23.75% ± 1.39% in session 3 versus 24.18% ± 0.57% for "expert" animals, Wilcoxon rank sum, not significant). The accompanying increase in the frequency of concatenated covert patterns in the rats' choices (Figure 2B, bottom) argues that the animals were indeed able to infer the underlying rule, rather than to achieve greater reward by simply biasing a stochastic strategy to select one port more frequently. In striking contrast, reward rates for rats that had previously been playing against competitor 3 rose to only, on average, 14.5% in three sessions, not above the rate for random behavior, implying that they did not discover the covert sequence (Figure 2B; not significant for competitor 3 group in sessions 3 versus 1; $p < 0.005$ in sessions 2 and 3 for competitors 2 versus 3, Wilcoxon rank sum). This was not due to a lack of rewarded examples, because in the first session these animals performed the covert sequence, presumably by chance, and received reward with a frequency on par with that of the animals pre-exposed to competitor 2 (frequency of "covert patterns" 10.36% ± 1.53% for competitor 2 group, 7.33% ± 1.13% for competitor 3 group, Wilcoxon rank sum, not significant).

In principle, the prolonged insensitivity to environmental feedback observed in the "covert pattern" task following exposure to competitor 3 could result from a behavioral strategy that is feedback dependent but tests a range of possible patterns too broad to discover the covert "LLR" sequence efficiently. To address this possibility, we simulated our experimental framework with the "animal" being represented by a deterministic reinforcement learning algorithm belonging to the widely used class of Q-learners (Sutton, 1990; Watkins and Dayan, 1992). The underlying algorithm estimates, through experience, the value of choosing either the left or the right port given the immediate history pattern of a particular length (from n = 1 to 6 steps in the past for the different Q-learners) and makes the choice that has the higher estimated value (Figure 2C, top; Experimental Procedures). The Q-learners were able to infer the covert pattern rule and achieve high rates of reward (comparable to animals in Competitor 2 group; Figure 2C, bottom) even when they needed to estimate and track the value of a large number of states. This indicates that the information given to animals was at least in principle sufficient to constrain even a large hypothesis space that they may be using to defeat competitor 3. Interestingly, despite varying the number of patterns tested by the Q-learners and exploring a large space of algorithm parameters (Experimental Procedures), we were unable to find Q-learners that performed as well against competitor 3 and as poorly on the "covert pattern" task as the rats in competitor 3 group without removing environmental feedback for the "covert pattern" task (Figure 2C, bottom). In summary, experiments and modeling show that

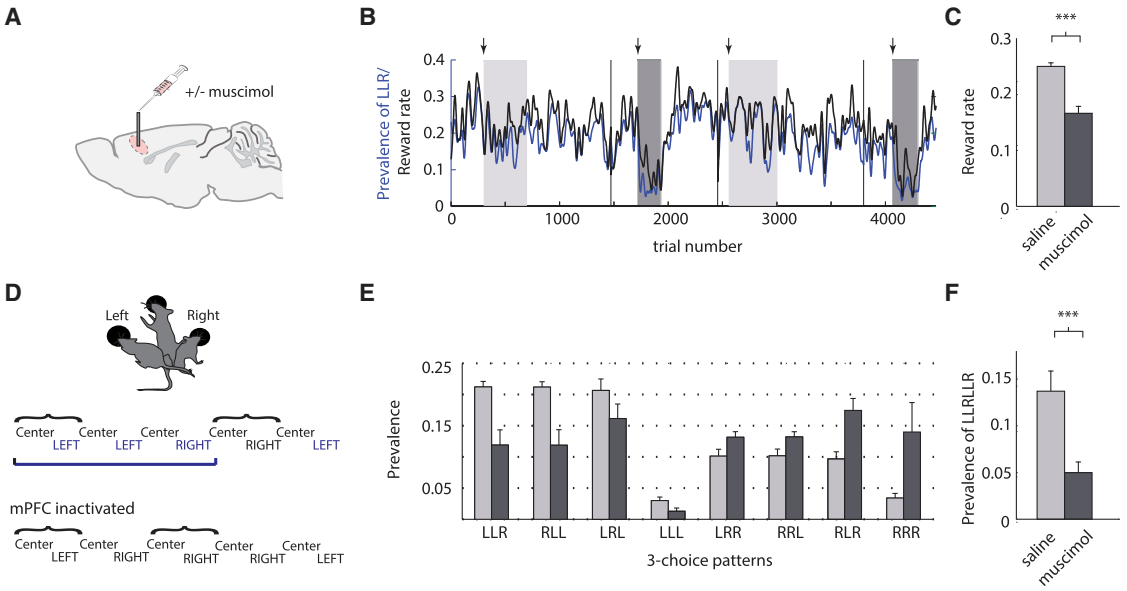

Figure 3. "Covert Pattern" Task Performance Is Dependent on ACC

(A) Injection of muscimol into ACC/mPFC (schematic).

(B) Concatenated and smoothed prevalence of LLR (blue line) and the reward rate (black line) during four consecutive sessions on different days (vertical lines indicate session boundaries) with vehicle (arrows indicate injections times; light gray bands indicate 2 hr intervals) or muscimol (arrows indicate injection times; dark gray bands indicate 2 hr intervals) injections.

(C) Mean reward rate for rats performing the covert sequence task during vehicle and muscimol application, respectively.

(D) Sequence of initiation port (center) and reward port (left or right) insertions with and without mPFC inactivation, top and bottom, respectively. Note that the initiation-to-reward-port transition (top curly brackets) is preserved. Bottom square bracket indicates covert sequence.

(E) Prevalence of all eight possible three-choice patterns during vehicle (light gray) and muscimol (dark gray) injections for the example shown in (B).

(F) Mean prevalences of two consecutive covert sequences during the 2 hr periods following vehicle (light gray) and muscimol (dark gray) injection, respectively.

(C and F) n = 5 animals. **p < 0.01, Wilcoxon rank sum. Error bars represent the SEM.

See also Figures S2 and S3.

animals lose their sensitivity to environmental feedback while playing against the strongest competitor (3), but not when playing against the intermediate competitor (2).

Suppression of ACC's Influence

Prior work suggests that feedback-dependent strategies adopted in competitive settings involve mental simulation of different possible outcomes (Abe and Lee, 2011; Abe et al., 2011; Zhu et al., 2012) and thus rely on an internal model that is thought to be stored, at least in part, in the ACC (Holroyd and Yeung, 2012; Karlsson et al., 2012; Matsumoto et al., 2003; Ribas-Fernandes et al., 2011; Yoshida and Ishii, 2006). To provide independent support for the conclusion that animals adopt a distinct behavioral mode against competitor 3 in which actions are not selected on the basis of past experience, we therefore investigated whether ACC still influences behavior during play against this competitor.

We first verified that effective performance on the "covert pattern" task does indeed rely on the ACC. We found that following a bilateral injection of muscimol, an agonist of the GABA-A inhibitory channel, into the broader area of medial prefrontal cortex (mPFC) (Figure 3A), the animals continued to make choices for hundreds of trials but no longer performed the covert sequence significantly above the value expected for a biased coin. The reward rate fell from 25.1% to 16.7% (averaged across all animals; Figure 3C, $p < 0.001$ for reward rate between saline and muscimol conditions, Wilcoxon rank sum, n = 8 animals). When we excluded from consideration animals for which the postmortem analysis showed that the injection was outside of the part of mPFC that is thought to be homologous to the ACC, the rate fell to 13% (example session in Figure 3B; Figure S2A). The observed decrease in reward rate following muscimol administration was due to a dramatic redistribution of the relative prevalence of higher order patterns in animals' choices (example for three-step patterns in Figure 3E; Figure 3F; $p < 0.001$ for "LLRLLR" between vehicle and muscimol conditions, Wilcoxon rank sum). Such a redistribution of patterns is evidence that following muscimol administration, past choices and outcomes have a reduced impact on current choices, as supported by Markov chain analysis (Figure S3; Experimental Procedures). ACC inactivation appeared to specifically affect complex sequencing rather than chaining of actions in general, because the animals still performed the sequential entries into the initiation and reward ports correctly (Figure 3D). Initiation port-reward port sequencing could, however, be reliably disrupted by injecting muscimol into the dorsomedial striatum instead of the ACC (Figure S2B). All muscimol effects were completely reversible, with performance returning to levels seen before the injection after approximately 2 hr, consistent with the duration of muscimol inactivation commonly observed (Martin, 1991). Thus, the ACC is

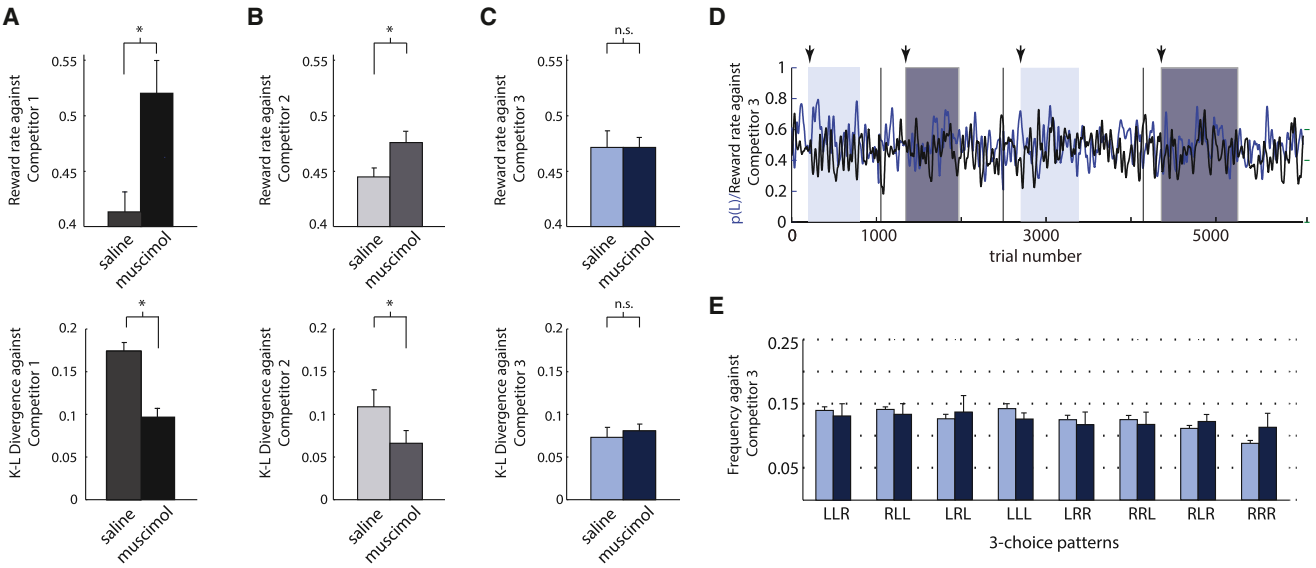

Figure 4. ACC Is Disengaged during Play against Competitor 3

(A–C) Mean reward rates (top) and mean Kullback-Leibler divergence of the rats' behavior, during play against Competitors 1–3 across 2 hr periods following vehicle (light bars) and muscimol (dark bars) injection, respectively.

(D) Prevalence of left choices (blue line) and the reward rate (black line), during four consecutive sessions on different days. Arrows indicate injection times for vehicle (light blue bands indicate 2 hr intervals) and muscimol (dark blue bands indicate 2 hr intervals).

(E) Frequencies of all eight possible three choice patterns during vehicle and muscimol injections for the example shown in (D). Example data are for the same animal as in Figures 3C and 3F after four sessions of retraining against competitor 3.

(A) n = 5 animals, (B) n = 4 animals, and (C) n = 6 animals. *p < 0.05, Wilcoxon rank sum. Error bars represent the SEM.

essential when strategic sequencing of actions is key to successful performance.

Next, we determined whether the strategy adopted against competitor 2 also depends on the ACC. When muscimol was administered into ACC as animals played against competitor 2, both the reward rate (Figure 4B, top, 44.48% ± 1.02% for saline versus 47.56% ± 0.81% for muscimol, $p < 0.02$, Wilcoxon signed rank, two tailed) and the behavioral variability (Figure 4B, bottom, K-L divergence decreasing from 0.109 ± 0.002 to 0.066 ± 0.015 under muscimol, $p < 0.04$, Wilcoxon signed rank, two tailed) increased significantly. The effect of ACC silencing was even greater when animals played competitor 1, consistent with the idea that the more structured behavior observed against this weaker competitor was also the result of counterprediction (Figure 4A). The strategy adopted against less sophisticated competitors thus requires ACC activity and presumably relies on computations that take place there, in line with primate and human studies (Abe and Lee, 2011; Barraclough et al., 2004; Zhu et al., 2012).

Does the strategy adopted against competitor 3 differ fundamentally from that used against less sophisticated virtual opponents? If it does, then the successful performance against competitor 3 might no longer rely on computations in ACC and thus might be unaffected by ACC inactivation. Indeed, neither reward rate nor, importantly, behavioral variability were affected when muscimol was administered while animals played against competitor 3 (Figure 4C). To guard against the possibility that the injection had missed ACC, we performed some of the perturbation experiments in animals where the effectiveness of the injection had already been established by its suppression of covert sequence performance prior to retraining against competitor 3 and observed the same dissociation (examples in Figures 3B and 3E; Figures 4D and 4E). Together with the dramatically reduced sensitivity to environmental feedback, this lack of any detectable effect of ACC inactivation when playing against the sophisticated opponent, strongly suggests that in this setting feedback- and model-dependent decision making is switched off. Our observations argue that the animals initially attempt to develop a more complex counterpredictive strategy as the competitive pressure increases, suggesting that a switch away from the "strategic" behavioral mode happens only when the search for a useful model is exhausted and deemed inadequate and thus when imposing variation on one's behavioral choices in a manner independent of prior beliefs and experience is most computationally advantageous.

Manipulation of LC Input into ACC Switches Behavioral Modes

Prior work has linked the discarding of an inadequate internal model with the rise in the level of noradrenergic neuronal activity (Jepma and Nieuwenhuis, 2011; Nassar et al., 2012), prompting us to examine the possibility that action of the noradrenergic system in ACC itself plays a key role in inducing the switch between behavioral modes. We achieved selective manipulation of noradrenergic terminals in ACC by targeting the expression of Channelrhodopsin 2 (ChR2) or DREADD receptors (Armbruster et al., 2007) to the noradrenergic neurons in the locus coeruleus of tyrosine hydroxylase-Cre (TH-Cre) transgenic rats (Witten et al., 2011) (Figure 5A) and then locally delivering light or the DREADD

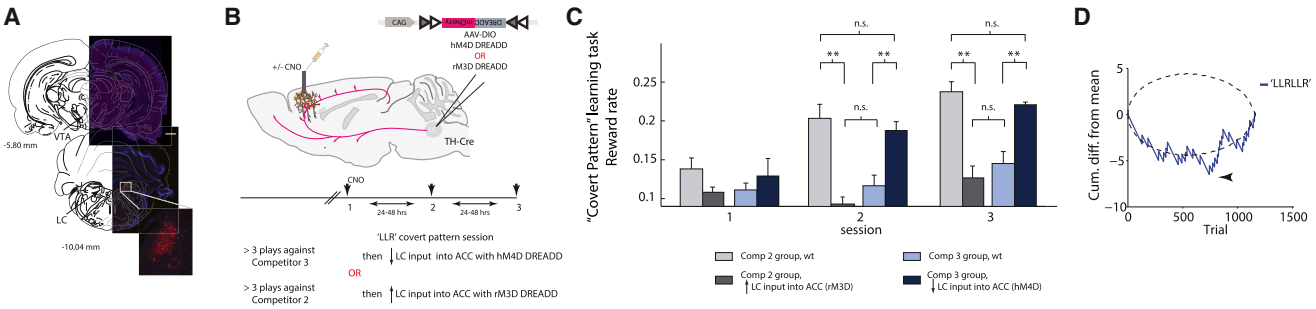

Figure 5. Manipulations of Locus Coeruleus Input into ACC Cause Switching between Behavioral Mode

(A) Specificity of local viral targeting strategy for effector delivery in TH-Cre rats. Expression of DIO-tdtomato virus in LC (bottom), but not in VTA (top).

(B) Top: experimental approach to pharmocogenetic control of LC input into the ACC (schematic). Bottom: experimental schedule.

(C) Reward rates in the three probe sessions following the switch to the "covert pattern" task for wild-type competitor 2 group (comp 2, light gray; data as in Figure 2B) and competitor 3 group (comp 3, light blue; data as in Figure 2B), as well as for CNO-treated LC-rM3D-DREADD animals pretrained against competitor 2 (dark gray, n = 5) or LC-hM4D-DREADD animals pretrained against competitor 3 (dark blue, n = 5 animals).

(D) Cumulative difference from the mean for the prevalence of "LLRLLR" concatenation for an example CNO session. The dashed line indicates the 95% confidence bound for the expected deviation. The arrow indicates a significant change point. ns, not significant, **p < 0.01, Wilcoxon rank sum. Error bars represent the SEM.

See also Figures S4, S5, and S6.

agonist Clozapine-N-Oxide (CNO) bilaterally in the ACC (Figures 5B, 6A, and 6B, left).

We first determined whether enhancing the action of the noradrenergic system in ACC triggers a switch away from the feedback- and model-dependent mode of action selection. Specifically, we asked whether a strong input from LC into the ACC would make competitor 2-exposed animals, which retained their sensitivity to environmental feedback, now behave on the "covert pattern" task more like animals whose modeling has been thwarted by the sophistication of competitor 3. To obtain pharmacological control over the activity of noradrenergic terminals, we relied on the rM3Ds DREADD receptor (Dong et al., 2010), which couples to the Gs-PKA signaling pathway implicated in facilitating neurotransmitter release (Maximov et al., 2007; Trudeau et al., 1996). Local administration of CNO into ACC of rM3Ds-expressing animals in the competitor 2 group prevented any performance improvement over the course of three learning sessions (Figure 5C), with reward rates that were indistinguishable from those seen in animals that had faced competitor 3 (9.30% ± 0.91%, 12.67% ± 1.53% in sessions 2 and 3 for rM3D competitor 2 group in the presence of CNO versus 11.65% ± 1.37%, 14.51% ± 1.54% for wild-type (WT) competitor 3 group, not significant, Wilcoxon rank sum). In striking contrast, when the learning experiment was later repeated under a vehicle condition in the same group of animals with a different three-step "covert pattern" ("RRL" instead of "LLR"), normal learning was observed (16.26% ± 1.25%, 21.05% ± 1.01% in sessions 2 and 3 for rM3D competitor 2 group in the presence of vehicle versus 20.33% ± 1.40%, 23.75% ± 1.30% for WT competitor 2 group, not significant, Wilcoxon rank sum; Figure S4A). Increased release from LC terminals in ACC thus appears to prevent animals from using their experience to infer or model the environment's governing rule and biasing their behavioral choices accordingly.

If activation of LC input into ACC indeed promotes the abandonment of the experience-derived internal model in favor of imposed behavioral variation, then suppressing it should restore the ability of animals to learn from environmental feedback. Next, we therefore examined whether animals whose modeling attempts had been thwarted by the sophistication of competitor 3 would regain their ability to discover the "covert pattern" efficiently if noradrenergic input into ACC was suppressed. To silence LC input into the ACC, we relied on a different DREADD receptor, hM4D, which couples to the Gi-GIRK pathway, thereby causing membrane hyperpolarization that inhibits action potential-triggered neurotransmitter release (Armbruster et al., 2007). Local administration of CNO into the ACC of hM4D -expressing animals that were operating in a feedback- and model-independent mode because of prior exposure to Competitor 3 lead to efficient learning in the "covert pattern" task (Figure 5C), with reward rates that were indistinguishable from those seen in animals that had faced competitor 2 rather than competitor 3 and had therefore retained sensitivity to environmental feedback (Figure 5C; reward rates for hM4D competitor 3 group in the presence of CNO: 18.83% ± 1.16%, 22.18% ± 0.33% for sessions 2 and 3, not significant from reward rates for WT Competitor 2 group: 20.33% ± 1.40%, 23.75% ± 1.30%, Wilcoxon rank sum). Thus, suppressing the action of the noradrenergic system in ACC appears to fully restore an animal's ability to learn from the environmental feedback.

Is the behavioral rescue selective for local noradrenergic action in ACC? The opposite behavioral consequence of ACC CNO administration in rM3D- versus hM4D-expressing animals (Figure 5C) strongly suggests that perturbation of the noradrenergic input rather than a nonspecific effect of the compound itself accounts for the observations. Furthermore, in line with what we observed for the wild-type animals after competitor 3 exposure, little learning over the course of three sessions was observed in the control group of hM4D-expressing animals that also previously faced competitor 3 but received either vehicle injection in or CNO injection outside of the ACC (reward rates for the LC-hM4D competitor 3 mixed control group: 11.86% ± 1.33%,

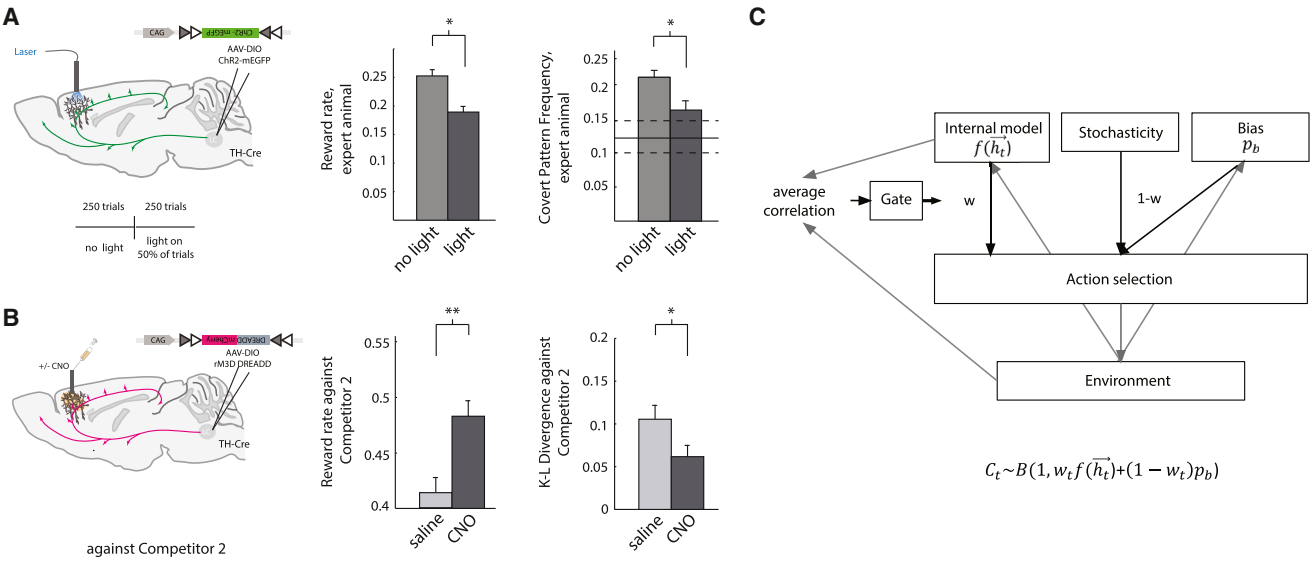

Figure 6. Enhancement of LC Input into ACC Leads to More Variable Behavioral Output

(A) Left, top: experimental approach to optogenetic enhancement of LC input into ACC (schematic). Bottom: experimental schedule. Middle: mean reward rate. Right: prevalence of LLR pattern, for an expert animal in the absence (light gray) and in the presence (dark gray) of ChR2-mediated enhancement of transmission from LC terminals in ACC. The solid and dashed gray lines in the right panel indicate mean and 95% confidence interval for the prevalence of LLR pattern expected for an unbiased stochastic process. n = 3 animals.

(B) Left: experimental approach to pharmacogenetic enhancement of LC input into ACC (schematic). Middle: mean reward rate. Right: mean K-L divergence of the rats' behavior, during play against competitor 2 under vehicle (light gray) and CNO (dark gray). n = 5 animals.

(C) Schematic of action selection. Behavioral choice on each trial is a single draw from a binomial distribution, where the probability of a left choice, p, is determined by a weighted combination of the recommendation of the internal model and a bias for or against the left option. The contribution of the model to the final choice is weighted by w, which decreases with increasing LC input into the ACC. *p < 0.05, **p < 0.01, Wilcoxon rank sum. Error bars represent the SEM.

15.25% ± 1.85% for sessions 2 and 3, not significantly different from reward rates in the WT competitor 3 group: 11.65% ± 1.37%, 14.51% ± 1.54%, Wilcoxon rank sum; Figure S4B). Finally, learning from feedback was not rescued by hM4D-mediated suppression of the input from the dopaminergic input from the ventral tegmental area (VTA) into the ACC (Figure S5). Combined, these control experiments argue for the selective role of input from the locus coeruleus in switching behavioral modes.

The fact that the CNO-treated hM4D animals show normal learning rates and achieve expert-level performance on a task that requires ACC (Figure 3) suggests that suppressing noradrenergic action in ACC restores model-based control of behavior. To obtain further support for this notion, we looked in greater detail at how the CNO-treated animals inferred the "left-left-right" rule. We specifically looked for a signature of hypothesis testing—an inference strategy whereby various discrete possible rules are tested until one is found that is consistent with the data. One of the most notable signs of hypothesis testing is the existence of abrupt change points in the learning curve where the frequency of the correct action pattern accelerates abruptly and, in particular, increases discontinuously during learning (Gallistel et al., 2004). Abrupt increases in the prevalence of single and concatenated "LLR" patterns were, in fact, seen for most animals in the group (Figures 5D and S6), suggesting that suppression of noradrenergic action in ACC restored not only the ability to learn from environmental feedback but also model-based control of behavior.

LC Input into ACC Triggers Behavioral Variation

Is the role of noradrenergic action in ACC limited to controlling learning in response to a sudden change in the environment as in the case of a switch to the "covert pattern" task? To address this question, we determined whether the influence of an animal's established model of the environment on behavioral choices can be suppressed by stimulating transmission from the LC terminals in the ACC.

We first assessed whether triggering release from LC terminals affects stable "expert" level of performance on the "covert pattern" task. We divided each behavioral session into two blocks of 250 trials during which LC input was either left unperturbed or enhanced through optical stimulation of ChR2-expressing terminals (Figure 6A). Illuminating ACC—which presumably caused the ChR2-mediated depolarization of and thus neurotransmitter release from LC terminals—led to a significant impairment in performance (with reward rate dropping from 25.27 ± 1.09 to 18.93 ± 1.02%, $p < 0.04$, Wilcoxon signed rank, two-tailed) which was caused by a drop in the prevalence of the "left-left-right" sequence (from 21.07% ± 1.09% in the absence to 16.27% ± 1.06% in the presence of stimulation, $p < 0.04$, Wilcoxon signed rank, two tailed; Figure 6A). Stimulating LC input into the ACC thus leads the animals to partially abandon a previously established behavioral model.

Finally, we determined whether enhancing LC input into the ACC prevents animals from using a counterpredictive strategy and makes them behave more randomly. Indeed,

rM3-DREADD-mediated enhancement of release from LC terminals in ACC during play against competitor 2 led to a significant a change in performance and in behavioral variability as did ACC inactivation (Figure 6B; c.f. Figure 4C; reward rate increasing from 41.42% ± 1.36% to 48.3% ± 1.39%, $p < 0.01$ and K-L-divergence dropping from 0.1055 ± 0.0162 to 0.0617 ± 0.0132, $p < 0.04$, Wilcoxon signed rank, two tailed). Combined, these results argue that the input from the locus coeruleus into the ACC controls the amount of imposed behavioral variation, whether or not learning is warranted by an environmental change.

DISCUSSION

The neural mechanisms responsible for variability in behavior are poorly understood. Prevailing views hold that in complex settings animals base their choice of actions on an inferred internal model of the environment's governing rules (Courville et al., 2006; Green et al., 2010; Nassar et al., 2010), with any behavioral variability attributed to noise (Faisal et al., 2008; Gold and Shadlen, 2007; Padoa-Schioppa, 2013) or errors in the inference of such rules (Beck et al., 2012). Here, we provide evidence that the brain possesses a mechanism for imposed behavioral variation and demonstrate that LC-mediated gating of neural activity in the ACC—the presumed locus of the animal's beliefs about the causal structure of its environment (Holroyd and Yeung, 2012; Karlsson et al., 2012; Matsumoto et al., 2003; Nassar et al., 2012; Ribas-Fernandes et al., 2011)—determines whether behavior is based on an experience-derived internal model or is varied independently of prior experience.

The highly variable choice selections made in the behavioral mode exposed by our strong competitor were dramatically less sensitive to environmental feedback. By generating such a high degree of behavioral variability while eliminating any simple relation to past experience, animals in this behavioral mode exhibit essentially stochastic action selection. It is unclear whether the failure to reach the reward rate of 50% expected from an unbiased stochastic process (mean reward rate against competitor 3 was ~48%, different from 50%, $p<0.001$, one sample t test, Figure 4C; data not shown) in this effectively stochastic mode is due to a small bias in an otherwise random process or derives from imperfections in the neural implementation of a pseudorandom generator. Regardless of whether this behavioral mode relies on a truly random process (which cannot be proven experimentally) or merely approximates it, the resulting choices would appropriately be captured by a stochastic exploration term in behavioral models.

What is the neural substrate of the behavioral variability observed when animals abandon the internal model in favor of this effectively stochastic choice? In principle, a switch in activity in the ACC itself could be the source of behavioral variability (Hayden et al., 2011). Alternatively, a circuit outside the ACC may actively introduce variability into a downstream decision circuit, in a manner analogous to the role imputed for the LMAN nucleus in song learning in the zebra finch (Fee and Goldberg, 2011; Kao et al., 2005). Finally, stochasticity could emerge in the decision circuit itself. Our finding that behavior in the stochastic mode is insensitive to the suppression of ACC activity suggests that variability is largely generated outside of the ACC and thus argues against the first scenario but is consistent with both the second and third scenarios. Removal of ACC input could magnify the effect of the external locus of variability or alternatively, in the absence of strong ACC input, a "winner-take-all" structure of decision circuitry could amplify small, internal noise-driven differences to generate choice variability (Wang, 2002).

Our analysis has focused on behavioral variability in an extreme scenario in which stochasticity could be uniquely advantageous. This, however, begs the question of how the transition to variable behavior occurs in more typical settings for which full stochasticity is not needed. Intuitively, the exploitation of knowledge accumulated through internal modeling needs to be counterbalanced by exploration designed to improve the model's accuracy and test its current validity (Cohen et al., 2007; Sutton and Barto, 1998). The degree of behavioral variability may thus need to be modulated according to subjects' uncertainty about their internal model of the environment. Our findings that model-based control of behavior is abandoned when LC input into the ACC is enhanced, but can be restored in the stochastic regime by lowering it, argues that the extent to which choices are informed by the internal model is dependent on modulation by the noradrenergic system (Figure 6C). In this context, recent observations from studies measuring pupillary responses in humans—a known consequence of LC activation—suggest that levels of noradrenergic signaling reflect the degree of uncertainty about the accuracy of one's internal model, with high levels associated with the discarding of an unreliable model (Nassar et al., 2012) and low levels linked to stabilization of an accurate model (O'Reilly et al., 2013). Because uncertainty about model reliability has been shown to translate into instability of ACC ensemble activity (Karlsson et al., 2012), LC input into the ACC may modulate—via norepinephrine itself or via other substances thought to be co-released by the noradrenergic fibers (Devoto et al., 2001; Xu et al., 1998)—the strength and/or coherence of ACC output. Modulating the effectiveness of ACC output in driving the downstream decision circuit could therefore translate the degree of the model accuracy into an appropriate balance between exploitation and exploration.

We note that complete abandonment of an internal model and adoption of a fully stochastic behavioral mode is normally maladaptive because of the associated insensitivity to new information. In rats, such a mode appears to be triggered when repeated modeling efforts prove to be ineffective and thus bears a similarity to the condition of learned helplessness thought to follow the sustained experience of the futility of one's actions. Intriguingly, functional imaging studies in humans have suggested that a chronic reduction in ACC activity might play a role in this disorder (Bauer et al., 2003), providing a potential mechanistic counterpart to the disengagement of ACC from the decision-making process that accompanies the switch into an effectively stochastic behavioral mode in rodents. The fact that the ability to discern environmental rules can be restored by suppressing the action of the noradrenergic system in ACC could pave a path

to a better understanding of and intervention in states of learned helplessness.

EXPERIMENTAL PROCEDURES

Subjects

Male Long Evans rats (300–450 g) were kept at 85% of their initial body weight before food restriction by providing them with 4–5 g food pellets a day. Experiments were conducted in accordance with the NIH guidelines for animal research and were approved by the Institutional Animal Care and Use Committee at Howard Hughes Medical Institute's Janelia Farm Research Campus.

Task Design

The virtual competitive setting was inspired by primate work (Barraclough et al., 2004; Lee et al., 2004). The computer was programmed to predict which reward port the animal would choose on the current trial. The prediction was made by using the history of the animal's performance up to that trial in the session.

Computer Competitor	Prediction Algorithm
Competitor 1	binomial test; reacts to large bias, similar to algorithm 2 in references (Barraclough et al., 2004; Lee et al., 2004)
Competitor 2	binomial test; reacts to any bias
Competitor 3	boosting with diverse features (Friedman et al., 2000)

For the "covert pattern" inference task, the computer rewarded every instance of a three-step "covert pattern," usually the "left-left-right" sequence. In addition, the animal was rewarded with 10% probability when it escaped prediction by competitor 2, which was running in parallel.

Simulated Play

During simulated play, the prediction algorithm used by a particular competitor, rather than playing against an animal, uses existing behavioral data from an individual animal having faced a different competitor. Data up to trial n are used to make a prediction of the animal's behavior at trial n+1. The simulated reward is determined using the same rules that govern real play; i.e., simulated reward accrued if the animal's behavior disagreed with the prediction.

Variability Metrics

Divergence from the Optimal Deterministic Strategy

Competitors 1 and 2 use conditional prevalences of the left and right choices given a particular history pattern of up to three steps in the past to inform their prediction. This implies that the optimal deterministic strategy is to keep track of every pattern up to that length and ensure that the conditional prevalence of going left or right is 0.5.

We quantified how different the observed behavior was from this optimal strategy by calculating the Kullback-Leibler divergence (D_{KL}) of the observed distribution of conditional prevalences given all patterns of lengths n = 1, 2, and 3 from the optimal one.

For each history pattern of choices and reward,

$$\vec{h},\ D_{KL\,\vec{h}} = \sum_{L,R} p\left(L \wedge R \middle| \vec{h}\right) \log_2 \frac{p\left(L \wedge R \middle| \vec{h}\right)}{0.5}.$$

For all patterns of combined choices and reward of length n,

$$D^n_{KL\ full} = \sum_{i=1}^{4^n} p\left(\vec{h_i}\right) D_{KL\,\vec{h_i}} - \frac{df}{1.3863\ length(session)},$$

where the last term corrects for the limited sample size. df stands for degrees of freedom, and in this case

$$df = \sum_{p\left(\vec{h}_i\right)>0} i - 1.$$

The final metric used to generate Figure 1E was

$$(\max) D_{KL} = \max_{n=1,2,3} D^n_{KL\ full}.$$

General Measure of Variability in the Observed Sequence of Choices

The degree of randomness in the sequences of animals' choices was characterized using Shannon entropy.

For all history pattern of length n (choices only), $\vec{h}_i$,

$$H = -\sum_{i=1}^{2^n} p\left(\vec{h}_i\right) \log_2 p\left(\vec{h}_i\right) - \frac{df}{1.3863\ length(session)}.$$

In this case,

$$df = \left(\sum_{p\left(\vec{h}_i\right)>0} i - 1 \right) n.$$

Multiple Linear Regression

The relationship between the past three trials and the rats' choices was examined by performing the following regression analysis:

$$C_t = \alpha + \sum_{n=1}^{3} \beta_n C_{t-n} + \sum_{m=1}^{3} \beta_{3+m} R_{t-n},$$

where C_t is choice (1 and −1 for left and right, respectively) on trial t, and R_t is reward (1 for rewarded, −1 for rewarded) on trial t. First-session data from each competitor group was used for the analysis, and the distribution of R^2 values across all of the rats was reported for each competitor.

Reinforcement Learning Model

Each Q-learner was parameterized by a learning rate, α, and a discount rate, γ, as specified by Equation 1:

$$Q_{t+1}(s_t, a_t) \leftarrow Q_t(s_t, a_t) + \alpha \left(R_t + \gamma \max_{a_{t+1}} (Q_t(s_{t+1}, a_{t+1})) - Q_t(s_t, a_t) \right), \quad \text{(Equation 1)}$$

where s is the state of the environment, a is the action, and R is the reward received.

The action that maximizes the sum of the estimated Q-value for potential future states and weighted exploration bonus, as specified by Equation 2, was chosen deterministically. The exploration bonus was calculated using the square root of the number of trials since the potential state occurred, ρ, weighted by a variable parameter, β_{rec}.

$$E_{val} = Q_t(s_{t+1}, a) + \beta_{rec} \sqrt{\rho} \quad \text{(Equation 2)}$$

Q-learners performed 1,000 trials against competitor 3, followed by 3,000 trials on the "covert pattern" task. For each memory length N (N = 1 to 6 for different Q-learners) and each tuple of the parameters α, γ, β_{rec} (varied between 0 and 1 in increments of 0.1), twenty simulations were run, and the performance was averaged. The performance against competitor 3 was summarized as the reward rate for the last 700 trials. Q-learners that achieved the reward rate of at least 40% against competitor 3 were used in simulations for the "covert pattern" task.

Markov Chain Analysis

The effect of the past choice patterns on the probability of the subsequent response was estimated by measuring how well a Markov chain of order n (for n = 1 to 5) fits the data.

$$\aleph^2 = \sum \frac{(f_{obs} - f_{est})^2}{\sigma^2},$$

where f_{obs} is the observed frequency for a given pattern of length n, f_{est} is the estimated frequency for the given pattern of length n based upon the observed frequencies of patterns of length n-1, and σ^2 is the variance of pattern frequencies across sessions.

Change Point Analysis

A cumulative sums bootstrap scheme (Hinkley, 1971) was used to detect the presence of a change point in the prevalence of a pattern of choices within a session during the "covert pattern" task. For each trial, we computed S_t, the cumulative sum of the difference from the mean prevalence, θ_t, of a particular choice pattern h:

$$S_0 = 0;\ S_t \sum_{j=1}^{t} \left(h - \theta_t\right)$$

95% confidence intervals were estimated by computing S_t for shuffled time series of the occurrence of h within the session. The presence of a significant change point was indicated by the deviation of the cumulative difference from the mean prevalence beyond these confidence intervals.

Perturbation Experiments

Muscimol (Tocris Bioscience; 0.50 μl of 0.1 μg/μl solution) or CNO (Enzo Life Sciences; 0.50 μl of 3 μM solution) was infused through a bilateral cannula that had been surgically implanted above the targeted brain region. All injections were done in awake animals, making it possible for the behavioral assay to resume immediately after injection.

Channelrhodposin Perturbation Experiments

On stimulation trials, a 1 s 10 Hz train of 50 ms pulses of 5 mW 473 nm light (Stratus 473-80, Vortran Technologies) was delivered, triggered by the detection of an initiation port entry, bilaterally through a fiber optic cannula.

Histology

To locate the tip of the injection needle, fluorescent beads or GFP-expressing adenoassociated virus were injected at the end of the final experimental session. Several days later, animals were euthanized, and brains were fixed with 4% paraformaldehyde and sectioned (50 μm coronal sections).

For further details about the Experimental Procedures, please refer to the Extended Experimental Procedures.

SUPPLEMENTAL INFORMATION

Supplemental Information includes Extended Experimental Procedures and six figures and can be found with this article online at http://dx.doi.org/10.1016/j.cell.2014.08.037.

AUTHOR CONTRIBUTIONS

D.G.R.T. and A.Y.K. designed the study. M.K. and K.B. designed competitor 3. D.G.R.T., A.Y.K., and M.P. performed experiments; M.M. implemented the reinforcement learning model; and D.G.R.T., A.Y.K., M.P., and M.M. analyzed the data. A.V. performed surgeries. D.G.R.T. and A.Y.K. wrote the manuscript.

ACKNOWLEDGMENTS

We are grateful to K. Deisseroth for providing the TH-Cre rat transgenic line and to A. Hantman for providing the hM4-DREADD virus. We thank H. Hou (Janelia Summer Undergraduate Program) for help with the exploratory set of behavioral experiments and K. Morris, S. Lindo, and M. Copeland for technical assistance with surgeries and histology. W. Denk, S. Druckmann, T.J. Florence, V. Jayaraman, T. Jessell, M. Karlsson, A. Leonardo, N. Spruston, and C. Zuker offered useful discussions and comments on the manuscript. K. Ris-Vicari created the task illustration. This work was supported by the Howard Hughes Medical Institute.

Received: June 30, 2014
Revised: August 22, 2014
Accepted: August 25, 2014
Published: September 25, 2014

REFERENCES

Abe, H., and Lee, D. (2011). Distributed coding of actual and hypothetical outcomes in the orbital and dorsolateral prefrontal cortex. Neuron *70*, 731–741.

Abe, H., Seo, H., and Lee, D. (2011). The prefrontal cortex and hybrid learning during iterative competitive games. Ann. N Y Acad. Sci. *1239*, 100–108.

Armbruster, B.N., Li, X., Pausch, M.H., Herlitze, S., and 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. USA *104*, 5163–5168.

Barraclough, D.J., Conroy, M.L., and Lee, D. (2004). Prefrontal cortex and decision making in a mixed-strategy game. Nat. Neurosci. *7*, 404–410.

Bauer, H., Pripfl, J., Lamm, C., Prainsack, C., and Taylor, N. (2003). Functional neuroanatomy of learned helplessness. Neuroimage *20*, 927–939.

Beck, J.M., Ma, W.J., Pitkow, X., Latham, P.E., and Pouget, A. (2012). Not noisy, just wrong: the role of suboptimal inference in behavioral variability. Neuron *74*, 30–39.

Buschman, T.J., Denovellis, E.L., Diogo, C., Bullock, D., and Miller, E.K. (2012). Synchronous oscillatory neural ensembles for rules in the prefrontal cortex. Neuron *76*, 838–846.

Cohen, J.D., McClure, S.M., and Yu, A.J. (2007). 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. Sci. *362*, 933–942.

Courville, A.C., Daw, N.D., and Touretzky, D.S. (2006). Bayesian theories of conditioning in a changing world. Trends Cogn. Sci. *10*, 294–300.

Daw, N.D., O'Doherty, J.P., Dayan, P., Seymour, B., and Dolan, R.J. (2006). Cortical substrates for exploratory decisions in humans. Nature *441*, 876–879.

Devoto, P., Flore, G., Pani, L., and Gessa, G.L. (2001). Evidence for co-release of noradrenaline and dopamine from noradrenergic neurons in the cerebral cortex. Mol. Psychiatry *6*, 657–664.

Dong, S., Allen, J.A., Farrell, M., and Roth, B.L. (2010). A chemical-genetic approach for precise spatio-temporal control of cellular signaling. Mol. Biosyst. *6*, 1376–1380.

Dorris, M.C., and Glimcher, P.W. (2004). Activity in posterior parietal cortex is correlated with the relative subjective desirability of action. Neuron *44*, 365–378.

Faisal, A.A., Selen, L.P., and Wolpert, D.M. (2008). Noise in the nervous system. Nat. Rev. Neurosci. *9*, 292–303.

Fee, M.S., and Goldberg, J.H. (2011). A hypothesis for basal ganglia-dependent reinforcement learning in the songbird. Neuroscience *198*, 152–170.

Friedman, J., Hastie, T., and Tibshirani, R. (2000). Additive logistic regression: a statistical view of boosting (with discussion and a rejoinder by the authors). Ann. Stat. *28*, 337–407.

Gallistel, C.R., Fairhurst, S., and Balsam, P. (2004). The learning curve: implications of a quantitative analysis. Proc. Natl. Acad. Sci. USA *101*, 13124–13131.

Gold, J.I., and Shadlen, M.N. (2007). The neural basis of decision making. Annu. Rev. Neurosci. *30*, 535–574.

Green, C.S., Benson, C., Kersten, D., and Schrater, P. (2010). Alterations in choice behavior by manipulations of world model. Proc. Natl. Acad. Sci. USA *107*, 16401–16406.

Hayden, B.Y., Pearson, J.M., and Platt, M.L. (2011). Neuronal basis of sequential foraging decisions in a patchy environment. Nat. Neurosci. *14*, 933–939.

Hinkley, D.V. (1971). Inference about the change-point from cumulative sum tests. Biometrika *58*, 509–523.

Holroyd, C.B., and Yeung, N. (2012). Motivation of extended behaviors by anterior cingulate cortex. Trends Cogn. Sci. *16*, 122–128.

Jepma, M., and Nieuwenhuis, S. (2011). Pupil diameter predicts changes in the exploration-exploitation trade-off: evidence for the adaptive gain theory. J. Cogn. Neurosci. *23*, 1587–1596.

Kao, M.H., Doupe, A.J., and Brainard, M.S. (2005). Contributions of an avian basal ganglia-forebrain circuit to real-time modulation of song. Nature *433*, 638–643.

Karlsson, M.P., Tervo, D.G., and Karpova, A.Y. (2012). Network resets in medial prefrontal cortex mark the onset of behavioral uncertainty. Science *338*, 135–139.

Lee, D., McGreevy, B.P., and Barraclough, D.J. (2005). Learning and decision making in monkeys during a rock-paper-scissors game. Brain Res. Cogn. Brain Res. *25*, 416–430.

Lee, D., Conroy, M.L., McGreevy, B.P., and Barraclough, D.J. (2004). Reinforcement learning and decision making in monkeys during a competitive game. Brain Res. Cogn. Brain Res. *22*, 45–58.

Martin, J.H. (1991). Autoradiographic estimation of the extent of reversible inactivation produced by microinjection of lidocaine and muscimol in the rat. Neurosci. Lett. *127*, 160–164.

Matsumoto, K., Suzuki, W., and Tanaka, K. (2003). Neuronal correlates of goal-based motor selection in the prefrontal cortex. Science *301*, 229–232.

Maximov, A., Shin, O.-H., Liu, X., and Südhof, T.C. (2007). Synaptotagmin-12, a synaptic vesicle phosphoprotein that modulates spontaneous neurotransmitter release. J. Cell Biol. *176*, 113–124.

Maynard Smith, J., and Harper, D.G. (1988). The evolution of aggression: can selection generate variability? Philos. Trans. R. Soc. Lond. B Biol. Sci. *319*, 557–570.

Nash, J.F. (1950). Equilibrium points in n-person games. Proc. Natl. Acad. Sci. USA *36*, 48–49.

Nassar, M.R., Wilson, R.C., Heasly, B., and Gold, J.I. (2010). An approximately Bayesian delta-rule model explains the dynamics of belief updating in a changing environment. J. Neurosci. *30*, 12366–12378.

Nassar, M.R., Rumsey, K.M., Wilson, R.C., Parikh, K., Heasly, B., and Gold, J.I. (2012). Rational regulation of learning dynamics by pupil-linked arousal systems. Nat. Neurosci. *15*, 1040–1046.

O'Reilly, J.X., Schüffelgen, U., Cuell, S.F., Behrens, T.E., Mars, R.B., and Rushworth, M.F. (2013). Dissociable effects of surprise and model update in parietal and anterior cingulate cortex. Proc. Natl. Acad. Sci. USA *110*, E3660–E3669.

Ölveczky, B.P., Andalman, A.S., and Fee, M.S. (2005). Vocal experimentation in the juvenile songbird requires a basal ganglia circuit. PLoS Biol. *3*, e153.

Padoa-Schioppa, C. (2013). Neuronal origins of choice variability in economic decisions. Neuron *80*, 1322–1336.

Page, S., and Neuringer, A. (1985). Variability is an operant. J. Exp. Psychol. Anim. Behav. Process. *11*, 429.

Payzan-LeNestour, E., Dunne, S., Bossaerts, P., and O'Doherty, J.P. (2013). The neural representation of unexpected uncertainty during value-based decision making. Neuron *79*, 191–201.

Ribas-Fernandes, J.J., Solway, A., Diuk, C., McGuire, J.T., Barto, A.G., Niv, Y., and Botvinick, M.M. (2011). A neural signature of hierarchical reinforcement learning. Neuron *71*, 370–379.

Sutton, R.S. (1990). Integrated architectures for learning, planning, and reacting based on approximating dynamic programming. Proceedings of the Seventh International Conference on Machine Learning *1990*, 216–224.

Sutton, R.S., and Barto, A.G. (1998). Reinforcement Learning: An Introduction-*Volume 1* (Cambridge: Cambridge University Press).

Trudeau, L.-E., Emery, D.G., and Haydon, P.G. (1996). Direct modulation of the secretory machinery underlies PKA-dependent synaptic facilitation in hippocampal neurons. Neuron *17*, 789–797.

Vickery, T.J., Chun, M.M., and Lee, D. (2011). Ubiquity and specificity of reinforcement signals throughout the human brain. Neuron *72*, 166–177.

Wang, X.-J. (2002). Probabilistic decision making by slow reverberation in cortical circuits. Neuron *36*, 955–968.

Watkins, C.J., and Dayan, P. (1992). Q-learning. Mach. Learn. *8*, 279–292.

Witten, I.B., Steinberg, E.E., Lee, S.Y., Davidson, T.J., Zalocusky, K.A., Brodsky, M., Yizhar, O., Cho, S.L., Gong, S., Ramakrishnan, C., et al. (2011). Recombinase-driver rat lines: tools, techniques, and optogenetic application to dopamine-mediated reinforcement. Neuron *72*, 721–733.

Xu, Z.Q.D., Shi, T.J.S., and Hökfelt, T. (1998). Galanin/GMAP- and NPY-like immunoreactivities in locus coeruleus and noradrenergic nerve terminals in the hippocampal formation and cortex with notes on the galanin-R1 and -R2 receptors. J. Comp. Neurol. *392*, 227–251.

Yoshida, W., and Ishii, S. (2006). Resolution of uncertainty in prefrontal cortex. Neuron *50*, 781–789.

Yu, A.J., and Dayan, P. (2005). Uncertainty, neuromodulation, and attention. Neuron *46*, 681–692.

Zhu, L., Mathewson, K.E., and Hsu, M. (2012). Dissociable neural representations of reinforcement and belief prediction errors underlie strategic learning. Proc. Natl. Acad. Sci. USA *109*, 1419–1424.