Neuron
Previews

When Brains Flip Coins

Thomas Akam[1,*] and Rui M. Costa[1,*]
[1]Champalimaud Neuroscience Programme, Champalimaud Center for the Unknown, Av. De Brasília, 1400-038 Lisbon, Portugal
*Correspondence: thomas.akam@neuro.fchampalimaud.org (T.A.), rui.costa@neuro.fchampalimaud.org (R.M.C.)
http://dx.doi.org/10.1016/j.neuron.2014.09.025

In a recent study in the journal *Cell*, Tervo et al. (2014) show that animals can implement stochastic choice policies in environments unfavorable to predictive strategies. The shift toward stochastic behavior was driven by noradrenergic signaling in the anterior cingulate cortex.

Adaptive behavior requires learning about predictable structure in the world in order to choose actions that maximize positive and minimize negative outcomes. A great deal of effort has been focused on understanding the diverse neural mechanisms that support learning about the expected values of future actions. However, in certain situations, slavishly following the recommendations of value learning systems may be counterproductive. One such situation arises in competitive interactions with other organisms, where deterministic behavior allows an opponent to predict your future actions, with consequences ranging from losing a game to becoming someone's prey. Another situation in which stochastic choice can be advantageous is in the tradeoff between exploration and exploitation. Exploitative choices toward the action currently believed to be best are taken at the expense of not exploring other options that, due to the changeable nature of the world, may now in fact be better. Exploration does not necessarily imply stochastic behavior; it may be preferable to deterministically guide exploratory choices toward those options about which there is the most uncertainty. However, while optimal exploration necessitates potentially complex computations of the informational payoffs of different choices, stochastic choice potentially offers a simple alternative for generating exploration. In the latest issue of *Cell*, Tervo et al. (2014) report experiments in which they challenged rats with a competitive game that drove them to exhibit stochastic behavior. Using genetically targeted manipulations, they identify a causal role in this stochasticity for a neuromodulatory system strongly implicated by prior work in controlling the balance between exploration and exploitation.

The competitive task used in the study required the rats to choose one of two reward ports on each trial. A computer competitor aimed to predict the rat's choice from their behavior and outcomes on previous trials. Reward was delivered only if the computer incorrectly predicted the animal's choice. The task is an adaptation of one previously used in primates (Barraclough et al., 2004), which in turn builds on work in the field of game theory in which the task is termed "matching pennies."

Animals were trained against virtual competitors of three different strengths. The weaker two used the animal's choice history to evaluate their choice bias following each unique sequence of actions and reward up to three trials in length. The strongest of these predictors was then used to guide the competitor's choice, with the weakest competitor only utilizing the prediction if the bias exceeded a certain level. The strongest competitor used a different approach, employing a machine learning method called boosting, which combined a large number of weak predictors, each based on different features of the history of prior actions and reward, to produce a more robust prediction of the animals' behavior. Tervo et al. (2014) present analyses indicating that the animals' behavior was more stochastic, and less dependent on the history of task events, when they played against stronger competitors.

Tervo et al. (2014) hypothesized that against the two weaker competitors the animals were employing a counterprediction strategy, leading to history dependence in their behavior, but that against the strongest competitor they switched to a stochastic, feedback-independent, behavior mode. To test this, Tervo et al. (2014) first trained animals against either competitor 2 (weaker) or competitor 3 (strongest) and then switched them to a new task in which a specific sequence of three choices (left, left, right) automatically led to reward. Rats previously trained against competitor 2 were able to find the covert sequence that reliably led to reward, while animals trained against competitor 3 appeared not to discover the covert sequences and received dramatically lower reward rates after three sessions. These striking effects could not be explained by a difference in how often the animals initially sampled the covert sequence, indicating a difference in learning between the groups and supporting the hypothesis that play against the strongest competitor caused the animals to switch to feedback-independent stochastic choice behavior.

Tervo et al. (2014) then inactivated a region of the dorsomedial prefrontal cortex called the anterior cingulate cortex (ACC), which has been implicated in reward-guided decision making but whose precise function remains contentious. ACC inactivation in animals playing the competitive task against competitors 1 and 2 caused their behavior to become more stochastic, which Tervo et al. (2014) interpret as evidence that ACC is necessary for counterprediction strategies. The authors did not observe any effect of inactivating ACC in animals playing against competitor 3, whose behavior was already highly stochastic. Inactivation of ACC during the covert sequence detection task severely impaired their ability to generate the rewarded sequence.

Tervo et al. (2014) proceeded to manipulate noradrenergic input into ACC from the locus coeruleus (LC) using a combination of pharmacogenetic and optogenetic approaches. Stimulating noradrenergic

(NA) input to the ACC in animals previously trained against competitor 2 severely impaired learning on the covert pattern task, recapitulating the effects observed in animals trained against competitor 3. Conversely, inhibiting noradrenergic input into ACC in animals trained against competitor 3 rescued their ability to learn the covert pattern task. These results indicate that increased noradrenergic input to ACC plays a mechanistic role in promoting the stochastic, feedback-independent behavior induced by play against competitor 3. Consistent with this, increased LC input to ACC during play against competitor 2 increased the stochasticity of the behavior. Elevated LC input into ACC was further shown to disrupt performance as well as learning of the covert pattern task using manipulations in extensively trained animals.

Interpretation of this striking but complex pattern of manipulation results necessarily depends on beliefs about how the animals are solving the tasks. As such, the computational problems posed by these tasks, and how they may be solved by the animals in the study, deserves close attention.

Solving the covert sequence task requires the ability to learn to take different actions dependent on the agent's recent action history. This is impossible for simple reinforcement learning (RL) agents typically used to model behavior on bandit style tasks, which treat the problem as one of choosing between two actions, left or right, with the choice on each trial made in the same state. One solution is the use of a richer world representation that treats choices preceded by different action histories as occurring in distinct states. Given such a state representation, the task can be solved by temporal difference methods, as demonstrated in the paper, in which the agent learns by trial and error the value for each action in each state. One interpretation of the effects of ACC inactivation on covert sequence task performance is therefore an inability to either form or use an appropriate history-dependent state representation. It is worth pointing out that though identifying an appropriate state representation can be seen as building a "model" of the world, solving the task does not require the use of a forward model that predicts future states given the current state and chosen action. As such, deficits in this task do not speak to whether animals are using model-based or model-free reinforcement learning as these terms are normally used (Sutton and Barto, 1998).

The covert pattern task can alternatively be solved by an agent lacking a history-dependent state representation but endowed with a richer repertoire of possible actions, which encompasses composite actions that are sequences of unitary left-right choices. The use of composite actions in learning, termed hierarchical RL, can offer striking advantages in complex environments (Botvinick et al., 2009). Given such an enriched action space, temporal difference methods are again sufficient to solve the task. The ACC inactivation results are thus consistent with recent proposals that ACC plays a role in hierarchical RL and specifically the selection of composite actions (Holroyd and Yeung, 2012).

In the matching pennies task, Tervo et al. (2014) interpret structure in behavior during play against weak opponents as evidence of counterprediction of the opponent's strategy. During play against the weakest opponent, some animals scored significantly above the 50% chance level, demonstrating that they are learning something useful. However, in matching pennies, but unlike in three action competitive games such as paper-scissors-stone, learning to counterpredict the opponent's strategy is hard to distinguish from learning action values through reinforcement learning. We also note that unlike in versions of matching pennies used in prior monkey experiments where visual stimuli indicated which of the two options the computer choose on each trial, the only feedback the animals received in the current study was the presence or absence of reward, i.e., they lack information about the counterfactual outcome that would have been available had they chosen the other option. Irrespective of whether the sensitivity to recent history observed in play against weak competitors is seen as evidence of counterprediction or reinforcement learning, performance above chance level in matching pennies almost certainly shares with the covert sequence task the requirement to learn about the value of actions following different choice histories.

The findings that enhanced NA input to ACC promotes stochastic behavior and reduces sensitivity to recent outcomes are consistent with two prominent theories of NA function. Yu and Dayan (2005) proposed that NA signals unexpected uncertainty, i.e., variability in the outcome of actions, above and beyond that predicted by recent observations. Unexpected uncertainty is a sign that something has changed in the environment and hence that previously learned predictive relationships are likely to be unreliable. The normative response to this lack of confidence in current beliefs is to reduce their influence over choice behavior, leading to increased stochasticity and exploration. Encoding of unexpected uncertainty by LC neurons has recently received support from human neuroimaging (Payzan-LeNestour et al., 2013). A largely compatible theory by Aston-Jones and Cohen (2005) proposed that enhanced tonic (as opposed to phasic) NA activity signals a shift to exploratory behavior, though unlike in Yu and Dayan's proposal this may occur in response to decreases in the utility of the previous behavior or due to evidence that the environment has recently changed. A recent study tracking pupil diameter, which is thought to correlate with baseline LC neuronal activity, provides correlational evidence for a role of NA signaling in exploratory choice (Jepma and Nieuwenhuis, 2011).

Though the NA manipulations identify a component of the mechanism through which playing against competitor 3 leads to a failure to learn the covert pattern task, these results raise many further questions. A key question is what is happening to activity in the LC during play against the strong competitor and subsequent failure to lean the covert pattern task, and specifically whether the LC has switched to the tonic state, with elevated baseline activity and reduced phasic responses, identified by Aston-Jones and Cohen with exploratory behavior. A second and related question is whether the use of DREADD receptors to manipulate release from NA terminals has differential effects on tonic and phasic responses. Assuming that the behavioral effects are indeed mediated

by a persistent change in LC activity, a next step is identifying the inputs to LC responsible for this change and the learning processes in these regions that cause the switch to stochastic behavior. If this switch is indeed a result of meta learning, i.e., learning about the extent to which lower-level controllers should be allowed to guide behavior, it is an interesting computational question why these same meta learning processes fail to return behavioral control back to value-based decision making in the face of the dramatic change in reward statistics when animals switch to the covert pattern task. Finally, the network dynamics that actually generate the stochastic choices remain to be identified and located. Clearly, these latest results are not the last we will hear from the stochastic side of the brains' behavioral repertoire.

REFERENCES

Aston-Jones, G., and Cohen, J.D. (2005). Annu. Rev. Neurosci. *28*, 403–450.

Barraclough, D.J., Conroy, M.L., and Lee, D. (2004). Nat. Neurosci. *7*, 404–410.

Botvinick, M.M., Niv, Y., and Barto, A.C. (2009). Cognition *113*, 262–280.

Holroyd, C.B., and Yeung, N. (2012). Trends Cogn. Sci. *16*, 122–128.

Jepma, M., and Nieuwenhuis, S. (2011). J. Cogn. Neurosci. *23*, 1587–1596.

Payzan-LeNestour, E., Dunne, S., Bossaerts, P., and O'Doherty, J.P. (2013). Neuron *79*, 191–201.

Sutton, R.S., and Barto, A.G. (1998). Reinforcement Learning: An Introduction. (Cambridge: MIT Press).

Tervo, D.G.R., Proskurin, M., Manakov, M., Kabra, M., Vollmer, A., Branson, K., and Karpova, A.Y. (2014). Cell *159*, 21–32.

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