Corticostriatal signatures of learning efficient internal models for control

Daniel McNamee (d.mcnamee@eng.cam.ac.uk)
Computational and Biological Learning Lab, University of Cambridge, Cambridge CB2 1PZ, UK.
Institute of Neurology, University College London, London WC1B 5EH, UK.

Matthew M. Botvinick (botvinick@google.com)
DeepMind, London N1C 4AG, UK.
Gatsby Computational Neuroscience Unit, University College London, London W1T 4JG, UK.

Samuel J. Gershman (gershman@fas.harvard.edu)
Department of Psychology and Center for Brain Science, Harvard University, Cambridge, MA 02138, USA.

Abstract

Control of high-dimensional, dynamical systems such as the body or the world imposes large complexity costs on the subserving neural hardware. We consider the hypothesis that, in order to make efficient use of its resources, the brain adaptively compresses its internal models via reinforcement learning. We study a normative measure of the importance of stimulus information in determining future action-outcome trajectories which can be updated via prediction errors. In a planning task, we found that decision reaction times were modulated by these predictions errors and that behavioral efficiency was strongly correlated with the strength of this modulation on a per-participant basis. Analysis of functional magnetic resonance imaging data indicated that three essential components of the model were encoded in focal cortical and striatal regions which were known to contain the necessary stimulus and action representations a priori. We suggest these data provide preliminary evidence that the brain monitors the efficiency of its internal models and updates them accordingly in the associative corticostriatal loop and that the results of these computations are reflected in sensorimotor loop activity and behavior.

Keywords: striatum, PFC, internal model, control, learning

Introduction

In instrumental conditioning, behavioral automaticity has been formalized into the notion of “habits”. Though much of the neural circuitry remains to be identified, there is evidence that dorsal areas of the striatum contains the stimulus-response associations which generate habitual actions (Balleine & O’Doherty, 2010). Many of the studies leading to this conclusion primarily consider responses consisting of single discrete actions such as lever-pressing. At the broader level of extended action sequences, neural signatures of “task brackets” have been recorded electrophysiologically in rodents engaged in a T-maze task (Smith & Graybiel, 2013). Comparatively, these spatiotemporal patterns of neural activity which appear to parse task behaviors into subcomponents are identified in dorsolateral striatum (DLS). However, the response characteristics of neurons in nearby dorsomedial regions (DMS) are also dynamically modulated as a function of learning and behavior in the same task paradigms (Kim, Lee, & Jung, 2013; Stalnaker, Calhoon, Ogawa, Roesch, & Schoenbaum, 2012; Thorn, Atallah, Howe, & Graybiel, 2010) and analogous tasks in humans (Tanaka, Balleine, & O’Doherty, 2008; McNamee, Liljeholm, Zika, & O’Doherty, 2015). The computational basis of the similarities and differences between neural activity in these two regions remains an open question (Smith & Graybiel, 2014). Anatomically, these regions are classified as components of two distinct corticostriatal loops with the DLS forming a key node in the “sensorimotor loop” and the DMS being part of the “associative loop”. However, the interconnections between these areas and relatively limited cytoarchitectonic divergence are suggestive of integrative functional contributions to control (Haber, 2016).

Here, in pursuit of a computational model of the functionalities of these regions in control, we develop and examine an algorithm for internal model compression via error-driven learning. The prediction errors in the model trade off internal model efficiency against the flexible and predictable control of the environment. The associated reinforcement learning algorithm converges to the *contingency information values* (see **Theory** for a precise definition) of stimuli such that stimuli with low estimated contingency information can be rationally excised from an internal model. This has multiple practical benefits in terms of behavioral efficiency. First, desired actions at such points in a control process can be pre-programmed. Second, it implies that the same action can be deployed regardless of variability in the precise stimulus input. Third, it enables sequences of actions which are interceded by such states to be judiciously concatenated and deployed as “open-loop” action sequences (Dezfouli & Balleine, 2013). On the other hand, states with high contingency information values must be protected from behavioral automation.

Methods

Theory

Related to recent information-theoretic approaches in artificial agents such as “empowerment” (Salge, Glackin, & Polani, 2013) and “intrinsic motivation” (Schmidhuber, 2010), we study a measure of the degree to which an agent should integrate a stimulus observation into its control processes which we refer to as *contingency information value*. Within a probabilistic framework of associative learning (Gershman & Niv, 2012), an internal model relates different states, S and O, to each other within a probabilistic graphical model which defined by conditional probability distributions such as $\mathrm{P}(O|S)$. Here, we consider an instrumental analogue of this formalism and thus incorporate actions A to form the distribution $\mathrm{P}(O|S,A)$ over future outcome states O. The uncertainty in the identity of the outcome state $o \in O$ is captured by the entropies $\mathbb{H}(O|S,A)$ and $\mathbb{H}(O|A)$ where the former depends on the information provided by both the states S and actions A while the latter is conditionally dependent on the actions alone. The difference between these two (equal to the conditional mutual information $I(O,S|A)$) measures the contribution of the identity of the intervening state S in determining future action-outcome contingencies. This is a measure $C(S)$ of the information contained in the specific identities of states $s \in S$ regarding future action-outcome trajectories:

$$\begin{aligned} C(S) &:= \mathbb{H}(O|A) - \mathbb{H}(O|S,A) \\ &\equiv I(O,S|A) \qquad (1) \\ &= \sum_{s,a,o} p(s,a,o)\left[\mathrm{L}(o|s,a) - \mathrm{L}(o|a)\right] \qquad (2) \end{aligned}$$

where $\mathrm{L}(x) := \log \mathrm{p}(x)$ denotes the negative description length of a value x of random variable X. The difference $\mathrm{L}(o|s,a) - \mathrm{L}(o|a)$ measures the amount information lost if only the prospective action variable a is used to predict subsequent states o (via $\mathrm{L}(o|a)$) compared to incorporating current state information s (as in $\mathrm{L}(o|s,a)$). High contingency information value implies that the identity of the intervening state $s \in S$ is important in determining the future action-outcome contingencies. If $C(S)$ is zero, the specific identity of the state $s \in S$ contains no contingency information regarding future action-outcome contingencies and thus can be eliminated in a compressed internal model.

We summarize the learning rules for updating estimates of a state-action-outcome log-probabilities which is used to model choice behavior as well as internal estimates of stimulus-specific contingency information values $\hat{C}_t(s)$. Let $\mathrm{L}(x) = \log \mathrm{p}(x)$ be the negative description length of a value x of random variable X. Based on a learning rate α, an estimate $\hat{\mathrm{L}}_t(s,a,o)$ of the generative log-probability $\mathrm{L}(s,a,o) = \log \mathrm{p}(s,a,o)$ can be updated after observing a trajectory $\boldsymbol{\tau} = (\boldsymbol{\tau}_s, \boldsymbol{\tau}_a, \boldsymbol{\tau}_o)$ as

$$\hat{\mathrm{L}}_{t+1}(s,a,o) = \hat{\mathrm{L}}_t(s,a,o) + \alpha\left[\mathrm{L}(\boldsymbol{\tau}|s,a,o) - \hat{\mathrm{L}}_t(s,a,o)\right] \qquad (3)$$

Since states are fully observable in the task studied here, the observation log-likelihood is $\mathrm{L}(\boldsymbol{\tau}|s,a,o) = \log(1)$ if $\boldsymbol{\tau} = (s,a,o)$ or else $\mathrm{L}(\boldsymbol{\tau}|s,a,o) = \log(0)$. Practically, we buffer $\mathrm{p}(\boldsymbol{\tau}|s,a,o)$ probabilities with very small values in order to avoid negative infinities in the log computations. We normalize these quantities at decision time via a softmax rule to predict participant decisions D given a target goal g:

$$\mathrm{P}(D_t = a|s,g) = \frac{e^{\beta \hat{\mathrm{L}}_t(s,a,g)}}{\sum_{g'} e^{\beta \hat{\mathrm{L}}_t(s,a,g')}} \qquad (4)$$

The learning rate α and choice "temperature" β were estimated from participant decisions for each participant and the model fit significantly better than a constant baseline choice model. Consistent with the previous learning process, an approximate updating of $\hat{C}_t(s)$ leads to the contingency information learning rule:

$$\begin{aligned} \hat{C}_t(s) &\approx \mathbb{E}_{a,o}\left[\mathrm{L}_t(o|s,a) - \mathrm{L}_t(o|a)\right] \\ \delta_{\mathsf{CPE}}(\boldsymbol{\tau}) &= \hat{\mathrm{L}}_{t+1}(s|a,o) - \hat{C}_t(s) \\ \hat{C}_{t+1}(s) &= \hat{C}_t(s) + \alpha\delta_{\mathsf{CPE}}(\boldsymbol{\tau}) \ . \qquad (5) \end{aligned}$$

We refer to $\hat{\mathrm{L}}_{t+1}(s|a,o)$ as a *credit assignment* signal which measures the degree to which a specific action-outcome can be attributed to a particular state s. If, on average, action-outcome combinations are not attributable to particular states, then discriminating between such states will not aid in predicting future outcomes contingent on action.

Task

We designed a challenging two-step task requiring flexible behavioral shifts in an environment with stochastic contingencies (Fig. 1). Participants were cued with one goal state $g \in O$ out of the four possible final outcomes (each associated with a particular color in a block of trials). The decision tree "branch" containing the goal color was indicated by the side of the screen that the goal cue was presented on. After a delay, participants responded with a finger or thumb button press $a \in A$ in order to select the branch. They were then randomly presented with one of two symbol cues $s \in S$ in each branch with equal probability of 0.5 (four distinct symbol cues over all). Each of the two outcomes $o \in O$ available in the branch were stochastically contingent on both actions with probabilities 0.2 and 0.8. Importantly, in the low contingency information condition, actions had the same outcome probabilities regardless of the symbol observed. In the high contingency information condition, the action-outcome contingencies reversed as a function of the symbols. Contingency information was manipulated within participants across six blocks of trials. In two blocks, the contingency information value in both branches was high, in another two, it was low in both, and in another two it was mixed. At the end of each trial, if the participant managed to arrive at the correct goal, then they were presented with a five dollar bill in addition to the outcome state cue. One randomly selected trial from each block was paid out to participants.

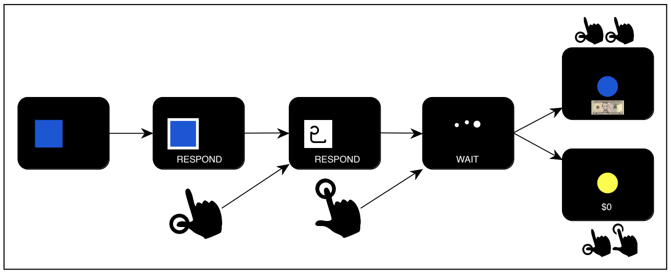

Figure 1: Time series of a single trial in the task.

Results

Behavior

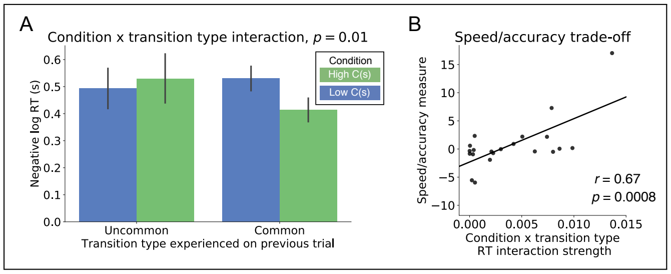

Figure 2: **A.** Bar plot of negative logarithm reaction times (RTs) as a function of condition and experienced transition type on the previous trial. Error bars reflect the standard deviations of RTs across all participants. Note that higher values indicate faster reaction times. **B.** The estimated interaction strengths are plotted as a function of a speed/accuracy trade-off measure on a per-participant basis. For the speed/accuracy trade-off measure, the ratio between the optimal choice rate and average reaction time was computed as a function of condition and then subtracted (Low $-$ High). Thus, this measure quantifies the degree to which participants' trade-offs improve in the low condition (where the intervening state can be rationally ignored) compared to the high condition.

Note that the task did not require any evaluation of stimulus contingency information in order to perform optimally. However, we observed (Fig. 2A) that the reaction time data reflected a significant interaction ($p = 0.01$) between the transition type experienced on the previous trial (whether the probability of the transition was common $\mathrm{P}(o|s,a) = 0.8$ or uncommon $\mathrm{P}(o|s,a) = 0.2$) and the contingency information condition.

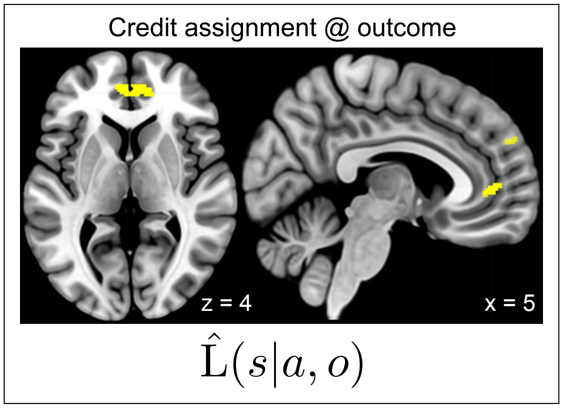

Figure 3: **Associative loop: mPFC/dACC.** Activity in medial prefrontal cortex (the ventral area of the dorsal anterior cingulate cortex) correlated with a credit assignment signal $\hat{\mathrm{L}}(s|a,o)$ at the time the outcome state was observed.

Neural Activity We summarize three sets of results from a "model-based" general linear model analysis of the functional magnetic resonance imaging data acquired while participants engaged in the task. All results are significant at $p < 0.05$SVFWE (family-wise error corrected in small volumes based on coordinates chosen from previous studies a priori (McNamee et al., 2015)) and are presented at $p < 0.005$ uncorrected. At the time of outcome presentation, activity in mPFC/dACC (Fig. 3) correlated with the "credit assignment" signal $\hat{\mathrm{L}}(s|a,o)$ necessary to compute the contingency information prediction error (see Eqns. 5). At the same timepoint, activity in dlPFC and hippocampus (Fig. 4) appeared to encode the prediction error itself δ_{CPE}. Shifting to the timepoint at which which the second action was performed (Fig. 5), neural activity in two key nodes in the sensorimotor loop, namely putamen and motor cortex, correlated with estimated contingency information $\hat{C}(s)$.

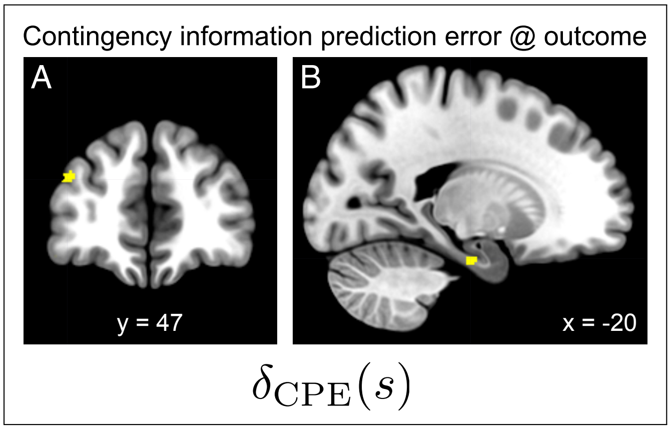
Contingency information prediction error @ outcome
A
B
y = 47
x = -20
$\delta_{\mathrm{CPE}}(s)$

Figure 4: **dlPFC and hippocampus.** Activity in dorsolateral prefrontal cortex and hippocampus correlated with the contingency information prediction error signal δ_{CPE} at the time the outcome state was observed. At the time of symbol appearance, activity in anterior caudate (not shown here) correlated with the decision variable $\hat{\mathrm{L}}(g|v,a)$ between state-action and goal.

Discussion

We presented analyses of neural activity which appear to encode signals from a model which learns, via prediction errors, if stimuli are necessary for predicting future action-outcome trajectories. "Credit assignment" signals in mPFC/dACC correlated with a representation of the degree to which an antecedent state was associated with an observed action-outcome contingency. This is a relatively complex calculation requiring information integration over multiple timepoints in the trial. It was established previously that this brain region predictively encodes the representations of state, action and outcome necessary to compute this quantity (McNamee et al., 2015). This signal is required to compute a contingency information prediction error which appeared to be represented in

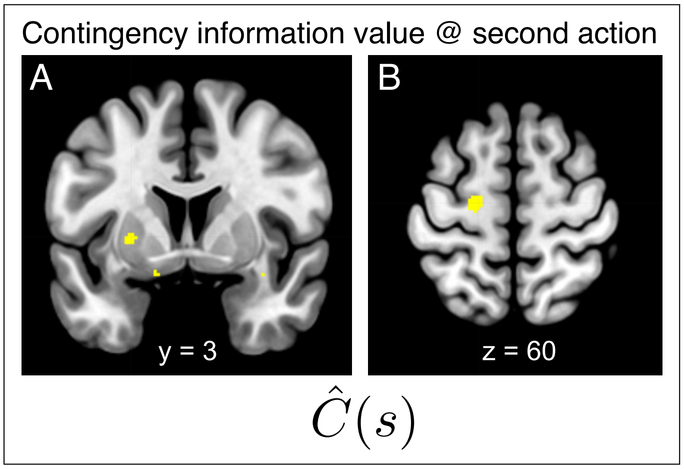

Figure 5: **Sensorimotor loop: DLS and motor cortex.** Activity in these areas correlated with estimated contingency information value $\hat{C}(s)$ at the time of the second response.

other nodes of the associative corticostriatal loop. The outputs of this learning process can make internal models and dependent action selection processes more efficient (Botvinick, Weinstein, Solway, & Barto, 2015). This was reflected in participant behavior and the prediction errors generated by this computation provide a mechanistic account of the observed reaction time effects (Fig. 2). Consider an uncommon transition in the high contingency information condition, this leads to a low credit assignment $\mathrm{L}(s|a,o)$ since (a,o) is much more likely to be observed in the alternative state. On average, this generates a negative contingency information prediction error $\delta_{\text{CPE}}(s)$ leading to a lower contingency information estimate $\hat{C}(s)$ and thus lower RTs subsequently. In contrast, the credit assignment measure for an uncommon transition in the low contingency information condition is higher since each state is equally unlikely to generate this transition and therefore this drives RTs up due to a large credit assignment to the observed state.

At the time of the response, activity in the sensorimotor loop correlated with the estimated state contingency information value. This is consistent with the transfer of an action representation to motor cortex which is contingent on the observed stimulus rather than automatically triggered by the previous action or pre-prepared during a planning phase. This process appeared to be modulated by the estimated importance of the stimulus in determining future action-outcome contingencies. Taken together, our preliminary results support the idea that the brain learns and utilizes efficient internal models during model-based control and that this learning process is driven by prediction error signals in the associative corticostriatal loop. This strengthens the evidence of a functional dichotomy between dorsolateral and dorsomedial striatum but integrates the respective roles of these regions within a common computational framework.

Acknowledgments

We thank the Wellcome Trust (Sir Henry Wellcome Fellowship 110257/Z/15/Z) and the National Institutes of Health (CRCNS 1R01MH109177) for funding.

References

Balleine, B., & O'Doherty, J. (2010). Human and rodent homologies in action control: corticostriatal determinants of goal-directed and habitual action. *Neuropsychopharmacology*, *35*(1), 48–69.

Botvinick, M., Weinstein, A., Solway, A., & Barto, A. (2015). Reinforcement learning, efficient coding, and the statistics of natural tasks. *Current Opinion in Behavioral Sciences*, *5*, 71–77.

Dezfouli, A., & Balleine, B. (2013). Actions, action sequences and habits: evidence that goal-directed and habitual action control are hierarchically organized. *PLoS Comput Biol*, *9*(12), e1003364.

Gershman, S. J., & Niv, Y. (2012). Exploring a latent cause theory of classical conditioning. *Learning & Behavior*, 1–40.

Haber, S. N. (2016). Corticostriatal circuitry. *Neuroscience in the 21st Century*, 1–21.

Kim, H., Lee, D., & Jung, M. W. (2013). Signals for previous goal choice persist in the dorsomedial, but not dorsolateral striatum of rats. *Journal of Neuroscience*, *33*, 52–63.

McNamee, D., Liljeholm, M., Zika, O., & O'Doherty, J. P. (2015). Characterizing the associative content of brain structures involved in habitual and goal-directed actions in humans: a multivariate fmri study. *Journal of Neuroscience*, *35*, 3764–3771.

Salge, C., Glackin, C., & Polani, D. (2013). Empowerment - an introduction.

Schmidhuber, J. (2010). Formal theory of creativity, fun, and intrinsic motivation (1990-2010). *IEEE Transactions on Autonomous Mental Development*, *2*, 230–247.

Smith, K. S., & Graybiel, A. M. (2013). A dual operator view of habitual behavior reflecting cortical and striatal dynamics. *Neuron*, *79*, 361–374.

Smith, K. S., & Graybiel, A. M. (2014). Investigating habits: strategies, technologies and models. *Frontiers in Behavioral Neuroscience*, *8*, 39.

Stalnaker, T. A., Calhoon, G. G., Ogawa, M., Roesch, M. R., & Schoenbaum, G. (2012). Reward prediction error signaling in posterior dorsomedial striatum is action specific. *Journal of Neuroscience*, *32*, 10296–10305.

Tanaka, S. C., Balleine, B. W., & O'Doherty, J. P. (2008). Calculating consequences: brain systems that encode the causal effects of actions. *Journal of Neuroscience*, *28*, 6750–6755.

Thorn, C. A., Atallah, H., Howe, M., & Graybiel, A. M. (2010). Differential dynamics of activity changes in dorsolateral and dorsomedial striatal loops during learning. *Neuron*, *66*, 781–795.