INTRODUCTION

The basal ganglia (BG) are connected to a large part of the cortex as part of cortico-basal ganglia-thalamo-cortical loops. What may be the role of the basal ganglia, given its prominent connectivity with the cortex? Initial theories regarding the role of these different loops described various parallel and independent circuits, each with a different function (Alexander, DeLong, Strick, & P. 1986). Each loop has been assumed to control a different type of movement, behaviour or cognitive process. Similarly, it has been suggested that the different loops are involved in distinct learning systems (Kim & Hikosaka, 2015; Redgrave et al., 2010; Yin & Knowlton, 2006). The associative or cognitive cortex–basal ganglia loop via the dorsomedial striatum (caudate nucleus in primates) has been associated with goal-directed behaviour, the selection of actions that lead to a desired outcome, while the sensorimotor cortex–basal ganglia loop that includes the dorsolateral striatum (putamen in primates) may be essential in habitual behaviour, ruled by stimulus-triggered actions. As both systems were hypothesized to operate independently of each other and could therefore propose different actions, an arbitration mechanism is required (Figure 1a). Arbitration may be implemented by means of the spiralling connections between the striatum and the midbrain dopaminergic neurons (Yin & Knowlton, 2006): the dorsomedial striatum could inhibit the activation of dopaminergic cells projecting to the dorsolateral striatum and thus render this loop silent. As an
alternative, Redgrave et al. (2010) proposed that arbitration may also be implemented directly in areas where the outputs of multiple loops converge, for example at the motor output neurons in the cortical or at the brainstem sensorimotor regions, but without providing details how this may work.

Pennartz, Ito, Verschure, Battaglia, and Robbins (2011) suggest that different cortex–basal ganglia loops rather use different inputs to predict outcomes, operate in parallel, but interact with each other in a ventral to dorsal direction. Joel and Weiner (1994) proposed that divergent connections within the basal ganglia emanating from the striatum, called split circuit, may establish automated responses; for example, a projection from the associative striatum to premotor GPi may lead to well-learned automatic motor responses. More recently, Dezfouli and Balleine (2012, 2013) and Yin (2014, 2016) proposed that the dorsomedial and dorsolateral loop may be hierarchically organized, where the goal-directed (dorsomedial) system represents a higher level of the hierarchy that controls the lower level habitual (dorsolateral) system. According to Dezfouli and Balleine (2012, 2013), the habitual system learns action sequences that can efficiently implement decisions made by the goal-directed system. In contrast, Yin (2014, 2016) claimed that the different loops are part of a negative feedback control system. In each loop, the difference between the desired value and the current state of an internal or external variable (reference signal) is computed within the striatum. This error signal is then integrated and sent as a reference signal to the next level of the hierarchy. Yin (2014, 2016) borrows ideas from control theory and assigns to the striatum the role of a comparator. For example, in the context of movement control the striatum receives a velocity reference and compares it to the current velocity also projected into the striatum. The output of the striatum represents a velocity error which is integrated by the SNr, and by integration, a position reference is obtained and sent to the next loop. This loop uses the same principle to compute a position coordinate error which can then be used to reach a desired location. Based on fMRI experiments using a task switching paradigm, Korb, Jiang, King, and Egner (2017) have also suggested a hierarchical organization where tasks are represented in the pre-supplementary motor loop and their response sets in the supplementary motor loop. No neuro-computational implementation of either of these frameworks has been reported so far. While we recognize a gradual shift from parallel and independent to hierarchical loops, the functionality of these loops is still under debate.

Compared to the control-inspired framework of Yin (2014, 2016), we propose that the striatum plays a critical role in learning rather than in online comparison of environment and reference signal. Further, we suggest that the
ultimate objective is broken down into a cascade of decisions that finally lead to an expected outcome. For example, in a navigation task, the ultimate objective could be to reach a particular location. Each of the cortex–basal ganglia loops represents a set of decision variables that are established by learning and each decision provides an objective that is sent to a loop of a lower hierarchy level. The task of the lower loop is then to determine decisions which will reach the objective provided by the higher-level loop. For example, the objective to reach a particular location provides a constraint to the election process in the lower level loop that determines particular actions to be taken to reach the location.

Hierarchical processing has been also proposed in non-habit learning tasks involving cortex–basal ganglia interactions such as in the context of task sets (Collins & Frank, 2013) and rule learning (Frank & Badre, 2012), or as general principles of prefrontal cortex organization (Badre & Nee, 2018; Koechlin, Ody, & Kouniheir, 2003; Nee & Brown, 2013).

As hierarchical processing is required to address multiple organizational principles, we introduce here a set of four key concepts (Figure 1b), integrated and evaluated by means of a neuro-computational model. According to the first key concept, each loop learns to select an intermediate objective at a different abstraction level, moving from goals in the ventral striatum via multiple cognitive and premotor loops to motor in the putamen. The advantage of such an organization is that a goal can be broken down into a subset of hierarchical decisions to be made. In each loop, the striatum learns to link combinations of objectives, sent by the previous loop, and contextual or sensory information to determine less abstract intermediate objectives. Thus, we agree with Pennartz et al. (2011) that loops process different inputs, but we emphasize the aspect of information transfer across loops. Such communication between loops is performed through overlapping cortico-striatal projections as supported by anatomical investigations by Haber (2016) and Groenewegen, Wouterlood, and Uylings (2017). Studies of the cortico-striatal projection in the rhesus monkeys have shown an overlap between 80% and 20% depending on the distance between cortical areas (Averbeck, Lehman, Jacobson, & Haber, 2014), similarly as observed in rats (Mailly, Aliane, Groenewegen, Haber, & Deniau, 2013).

The second key concept posits that the cortex integrates the objective selected by the basal ganglia with environmental information regarding the achieved objective. Cortical cells become initially activated by the selected objective but then follow the achieved result (regardless of the previous selection). Thus, the basal ganglia learn from the actual result of the decision and not from the expected result, a crucial aspect for solving the credit assignment problem.

The third key concept proposes shortcuts between loops. We have previously shown how a pathway that bypasses the basal ganglia can explain the effect of pallidotomy in parkinsonian patients, which after surgery are able to maintain previously acquired associations but cannot learn new ones (Baladron & Hamker, 2015; Schroll, Vitay, & Hamker, 2014). Further, the basal ganglia have been proposed to train meaningful cortico-cortical connections (Askey, Turner, & Horvitz, 2010; Collins & Frank, 2013; Villagrasa et al., 2018). In extension to these ideas, we will show that a shortcut between loops, when correctly trained, can directly link stimulus information with the final stages of the hierarchical process, ignoring the influence of goal representations, providing the motor loops less cognitive influence and leading to habitual and faster behaviour. Any disruption of the shortcut should however return control to the loops. Recent experiments indicate that the infralimbic cortex (IL), a part of the ventromedial prefrontal cortex, may be a possible cortical site for such a shortcut, as it plays a crucial role in the development of habitual behaviour. A lesion of the IL before training impedes the development of habitual behaviour (Killcross & Coutureau, 2003) while lesions after extended training change the behaviour of rats back to goal-directed actions after expressing habitual associations (Coutureau & Killcross, 2003; Smith, Virkud, Deiseroth, & Graybiel, 2012). Additionally, data show that habit-related activity develops first in the dorsomedial striatum and then, only after overtraining, in the IL (Smith & Graybiel, 2013). This suggests that the IL is part of a slower learning process that may be modulated by the basal ganglia.

The fourth key concept suggests that each cortex–basal ganglia loop computes its own prediction error signal used for learning. While the most common approach to learning in the basal ganglia relies on a reward prediction error signal for learning. While the most common approach to learning in the basal ganglia relies on a reward prediction error signal for learning. While the most common approach to learning in the basal ganglia relies on a reward prediction error signal for learning. While the most common approach to learning in the basal ganglia relies on a reward prediction error signal for learning. While the most common approach to learning in the basal ganglia relies on a reward prediction error signal for learning. While the most common approach to learning in the basal ganglia relies on a reward prediction error signal for learning. While the most common approach to learning in the basal ganglia relies on a reward prediction error signal for learning. While the most common approach to learning in the basal ganglia relies on a reward prediction error signal for learning. While the most common approach to learning in the basal ganglia relies on a reward prediction error signal for learning.
We have tested a neuro-computational implementation of the four concepts outlined above. The model comprises two cortex–basal ganglia loops and is tested on two cognitive tasks used to measure the development of habitual behaviour (Packard & McGaugh, 1996; Smith & Graybiel, 2013). Our contribution here is to explore the transfer of information from goals to actions via objectives, particularly with respect to the role of the four key concepts. We first ground our neuro-computational model in behavioural electrophysiological data. Our numerical simulations show that the model can account for the effects of devaluation, in the task of Smith and Graybiel (2013), and lesions to the caudate nucleus, in the task of Packard and McGaugh (1996). We further demonstrate the functionality of learning shortcuts for habitual behaviour and compare neural activation of the shortcut in the model with data from single-cell recordings in the IL.

We then address in detail the functional role of the key concepts by comparing different model designs on a hypothetical task, but similar to the one of Smith and Graybiel (2013), particularly with respect to outstanding questions such as the transfer of knowledge across tasks. As non-hierarchical models of decision-making have to associate a state to an action, they critically suffer from changes in the environment. We provide evidence that our outlined key concepts facilitate the transfer of knowledge across tasks as more abstract intermediate representations generalize to different goals. Finally, we outline an experimental design to further experimentally validate our theory.

2 | MATERIALS AND METHODS

2.1 | Modelling framework

Our framework suggests that multiple cortico-basal ganglia-thalamo-cortical loops are hierarchically organized and cooperate to generate behaviour (Figure 1b). Each loop selects an objective for the next loop in the hierarchy. The selected objective is initially stored in the cortex but also forwarded to the striatum of the next loop by overlapping cortico-striatal projections. The striatum of the next loop integrates this objective with environmental information, that is preprocessed sensor input to determine a decision and so on. The integration in the striatum is established by dopamine-modulated Hebbian learning. We assume that dopaminergic neurons in the midbrain compute a prediction error signal that is sent to the basal ganglia. As common in previous models, dopamine is then used as a third factor in Hebbian learning and thus determines the critical periods of learning. While the concept of a reward prediction error is well established, we propose non-uniform dopamine signals across the loops, and extend it by loop inherent prediction errors, indicated in Figure 1b as DA1, DA2 and so on.

We will demonstrate that one of the advantages of this particular hierarchical structure is the possibility of learning even in the absence of reward using information about the environment. In short, cortical cells encode the potential objectives from which each loop can select (e.g. dependent on the involved loop these cells may code positions in the environment or particular movements). As these neurons do become active not only once the basal ganglia have made a selection (planned objective), but also when the objective they represent has been achieved (“reached objective” in Figure 1b), dopamine-modulated Hebbian learning can initiate changes in the connections from the neurons encoding the achieved objective to the active striatal cells. Accordingly, learning is not limited to the evaluation of the current selection, but every action can lead to novel associations.

Additionally, we propose the existence of shortcuts which can partially bypass loops. These alternative cortico-thalamo-cortical pathways (Sherman & Guillery, 2011) use the thalamus as a relay to directly link sensory information to intermediate objectives or connect a loop with another further ahead. Similar to other cortico-cortical connections (Villagrasa et al., 2018), shortcuts can be trained and monitored by the BG through its output projections. Selection is initially fully controlled by the basal ganglia but may become strongly biased by the shortcut if appropriately trained.

We have created a neuro-computational implementation of the concepts outlined above composed of two basal ganglia loops (dorsomedial and dorsolateral network) and one shortcut (IL) that bypasses the first loop. The dorsomedial striatum receives a signal indicating a desired goal, for example, to obtain a particular juice reward. For simplicity, this goal selection process is not explicitly modelled. Goal selection at this level is likely performed by the limbic network involving the frontal cortex, the amygdala, the hypothalamus, the ventral striatum and the hippocampus (Balleine, Killcross, & Dickinson, 2003; Corbit, Muir, & Balleine, 2001; Gonner, Vitay, & Hamker, 2017; Groenewegen, Wright, Beijer, & Voorn, 1999; Groenewegen, Wright, & Uylings, 1997) based on needs and motivation. Further, the striatum receives sensory signals providing information about the present environmental (and internal) state. During learning, striatal neurons will become selective to combinations of sensory and goal signals. These striatal representations will be linked via the GPi and the thalamus to cortical cells of the loop which encode desired states of the environment.

The desired states in turn provide a reference signal for the dorsolateral loop. Thus, the dorsolateral striatum links desired states (and equally, the reached state) via a divergent projection from dorsomedial cortex with additional contextual information. The latter represents additional information required to select the appropriate action in order to reach the intended state selected by the previous loop. Learning will make cells selective to combinations of a desired state and contextual information and link them to the output cells which represent possible responses.

As outlined above, the dopamine signal of each loop differs (Figure 1b). In order to avoid wrong associations, the
dopamine signal should carry the relevant prediction error information for the content to be learned in each loop. Therefore, a different population of dopamine cells projects to each loop. In the first loop of our computational implementation, dopamine encodes a classical reward prediction error signal (Schultz, 2010). However, in the dorsolateral loop we compute an action consequence prediction error. Thus, a rise in the phasic dopamine level will occur only if the state reached after executing an action is different from the predicted one. In order to compute the prediction part of the prediction error signal, each dopamine population receives inhibitory projections from the striatum D1 neurons of the corresponding loop (Figure 2). Ongoing learning leads to a selective increase of this projection and thus reduces the activation of the dopaminergic cells. We will compare different models, a uniform (reward) prediction error signal in both loops with different prediction error signals and models with or without a shortcut.

2.2 Model details

In each of the two loops involving the BG, the direct pathway (striatum–internal globus pallidus–thalamus) learns to select the appropriate states, the short indirect (striatum–external globus pallidus–internal globus pallidus–thalamus) to inhibit common mistakes and the hyperdirect (subthalamic nucleus–internal globus pallidus–thalamus) to introduce surround inhibition by exciting incorrect states (Baladron & Hamker, 2015; Schroll et al., 2014). Dopamine-modulated Hebbian learning shapes both cortico-striatal and striato-pallidal projections (Schroll et al., 2014; Villagrasa et al., 2018). Although plasticity in the GP is not common in BG models and this property is not essential for the results reported here, we designed each loop dependent on our previously developed BG model. In short, plasticity in the GP has the advantage that striatal cells are not required to be hard wired to a single action or objective, consistent with data of category learning (Villagrasa et al., 2018). It is known that dopamine neurons do not only innervate the striatum, but also other nuclei of the BG (Schroll & Hamker, 2013) and administration of the dopamine precursor levodopa (Prescott et al., 2009) or high-frequency stimulation (Milosevic et al., 2019) has been shown to affect synaptic plasticity in SNr/GPi. Figure 2 shows a diagram of the BG part of our model.

Each nucleus contains a predefined number of rate-coded neurons. The striatum is further divided into two groups, one for D1 dopamine receptor cells and one for D2 dopamine receptor expressing cells. The membrane potential (mj) and firing rate (rj) of each cell is computed using the following equations:

$$
\tau \frac{dm_j}{dt} + m_j = \sum_{i \in N_e} w_{ij} r_i - \sum_{i \in N_i} w_{ij} r_i + B + \epsilon_j
$$

(1)

where r is a time constant, wij the weight from presynaptic neuron i to postsynaptic neuron j, Ne is the group of cells with an excitatory synapse to cell j, Ni is the group of cells with an inhibitory synapse to cell j, B is a baseline whose value depends on the nucleus and \(\epsilon_j\) is a noise term drawn from a uniform distribution. (\(j^+\)) converts negative numbers to 0. Parameter values for each nucleus are presented in Table 1.

Additionally, the striatum of both loops includes a set of cells through which direct cortical feedback is implemented. These neurons have afferent connections from the cortical cells of their corresponding loop and project to the pallidum in order to enhance learning in the late stages of each pathway. A similar mechanism was used in the original model (Schroll et al., 2014) to implement thalamic feedback.

Learning in the projections to both the dorsomedial and dorsolateral striatum follows the three-factor learning rule:

$$
\tau_w \frac{dw_{ij}}{dt} = f_{DA} \left( DA(t) - B_{DA} \right) C_{ij} - \alpha_j \left( r_j - r_{POST} \right)^2
$$

(2)

where wij is the weight between presynaptic neuron i and postsynaptic neuron j, fDA(\(DA(t) - B_{DA}\)) represents the effect of phasic changes of the dopamine level DA(t) with respect to its baseline value \(B_{DA}\). C corresponds to the correlation between the activity |

---

**FIGURE 2** Diagram of our basal ganglia circuit based on Schroll et al. (2014). Each loop in the model includes a direct pathway (striatum D1–GPi–thalamus), a short indirect pathway (striatum D2–GPe–GPi–thalamus) and a hyperdirect pathway (STN–GPi–thalamus). For simplicity, the cortical feedback pathway is not displayed.
of the pre- and postsynaptic cells and $\alpha_j \left( r_j - \bar{r}_{POST} \right)^2$ is a normalization term which limits the weight increase. $\bar{r}_{POST}$ is the mean activity of the postsynaptic population.

As motivated by recent data (Shen, Flajolet, Greengard, & Surmeier, 2008; Villagrasa et al., 2018), the effect of dopamine on plasticity is different on projections to striatal D1 and D2 cells. A phasic dopamine increase supports long-term potentiation in the projections to striatal D1 cells and long-term depression in the projections to striatal D2 cells. A phasic decrease of dopamine will have the opposite effect. This is controlled by the parameter $T_d$, which is set to 1 in connections to D1 cells and to $−1$ in projections to D2 cells. The final form of the equation for $f_{DA}$ is then:

$$f_{DA}(x) = \begin{cases} 
(T_d \cdot K_b \cdot x) & \text{if } (T_d \cdot x) > 0 \\
(T_d \cdot K_d \cdot x) & \text{if } (T_d \cdot x) < 0 \cup T_C \cdot C > 0 \\
0 & \text{else} 
\end{cases}$$

(3)

where $T_d$ controls the sign of the dopamine modulation. $K_b$ and $K_d$ control the speed of plasticity. $K_b$ is 1.6 for projections to the dorsomedial striatum and 2.4 for those to the dorsolateral striatum, and $K_d$ is 0.1 for all projections. $K_b$ is much larger than $K_d$ to ensure stronger LTP than LTD, which is functionally important but also consistent with data (Shen et al., 2008).

The correlation measure $C_{ij}$ ensures that plasticity affects only active postsynaptic cells. The equation for the dorsomedial striatum is:

$$C_{ij} = \left( r_i - \bar{r}_{PRE} - \gamma_{PRE} \right) \left( r_j - \bar{r}_{POST} - \gamma_{POST} \right)^+$

(4)

where $r_i$ is the firing rate of the postsynaptic cell and $r_j$ of the presynaptic cell, $\bar{r}_{POST}$ is the mean ring rate of the postsynaptic cell population and $\bar{r}_{PRE}$ is that of the presynaptic cell population, and $\gamma_{PRE}$ is a threshold for the presynaptic activity and $\gamma_{POST}$ is a threshold for the postsynaptic activity.

The correlation measure is inspired by the covariance learning rule, which is more powerful than basic Hebbian learning, when the input signals are not normalized to zero mean (Dayan & Abbot, 2001). In covariance learning, a threshold is subtracted from the rate of the neuron. This threshold should not be fixed, but adjusted to the pre- or postsynaptic activity. Covariance learning with a temporal mean has been demonstrated to be very biologically plausible (Fregnac et al., 2010), including mono-synaptic LTD. However, for large pools of neurons, temporal means are very unstable and inaccurate mean subtraction can severely affect learning performance (Loewenstein, 2008). Population means, however, have been demonstrated to lead to very stable learning results (Cossell et al., 2015; Wilschut & Hamker, 2009). While the population mean is in this strict way biologically not plausible, it could be considered as an approximation of the role of the inhibitory network in learning, which collects activation over a broader set of neurons.

In the dorsolateral loop, a trace ($\bar{r}_j$) of the firing rate is required. This is because each time a final state is reached, the input to the loop changes, modifying therefore the activity of the pre- and postsynaptic cells and $\alpha_j \left( r_j - \bar{r}_{POST} \right)^2$ is a normalization term which limits the weight increase. $\bar{r}_{POST}$ is the mean activity of the postsynaptic population.

| Parameter | Dorsomedial loop | Dorsolateral loop |
|-----------|------------------|------------------|
| Striatum D1 | | |
| Number of cells | 8 | 4 |
| $\tau$ | 10 | 10 |
| $\varepsilon_j$ | 0.3 | 0.05 |
| $B$ | 0 | 0 |
| Striatum D2 | | |
| Number of cells | 8 | 4 |
| $\tau$ | 10 | 10 |
| $\varepsilon_j$ | 0.01 | 0.01 |
| $B$ | 0 | 0 |
| STN | | |
| Number of cells | 8 | 4 |
| $\tau$ | 10 | 10 |
| $\varepsilon_j$ | 0.01 | 0.01 |
| $B$ | 0 | 0 |
| GPe | | |
| Number of cells | 2 | 2 |
| $\tau$ | 10 | 10 |
| $\varepsilon_j$ | 0.05 | 0.001 |
| $B$ | 1.0 | 1.0 |
| GPi | | |
| Number of cells | 2 | 2 |
| $\tau$ | 10 | 5 |
| $\varepsilon_j$ | 0.3 | 0.005 |
| $B$ | 1.5 | 1.1 |
| Thalamus | | |
| Number of cells | 2 | 2 |
| $\tau$ | 10 | 8 |
| $\varepsilon_j$ | 0.0 | 0.0 (0.425, 0.1) |
| $B$ | 0.0 | 0.0 (0.5) |
| Cortex | | |
| Number of cells | 2 | 2 |
| $\tau$ | 40 | 30 |
| $\varepsilon_j$ | 0.01 | 0.01 |
| $B$ | 0 | 0 |
pattern that produced the selected action. The traces allow the model to maintain a memory of the activations that really produced the selection. The equations for $C_{ij}$ in the dorsolateral loop are then:

$$C_{ij} = \left( r_i - \bar{r}_{PRE} - \gamma_{PRE} \right) \left( \bar{r}_j - \bar{r}_{POST} - \gamma_{POST} \right)$$

$$\frac{dTr_j}{\tau_r} = r_j - Tr_j$$

The factor $\alpha_j$ in the normalization term of Equation 2 is adaptive. Normalization becomes effective only if the activity of the postsynaptic cell is larger than a fixed threshold $m_{MAX}$. The temporal evolution of $\alpha_j$ is then given by:

$$\frac{d\alpha_j}{dt} + \alpha_j = (r_j - m_{MAX})$$

The parameters for each projection from the cortex to the BG are presented in Table 2.

Learning in the projections from the striatum to the pallidum happens between active cells in the striatum and pallidal cells whose activity is below the mean and not above possible final states (east and west ends of the maze). This is necessary because pallidal neurons have a high baseline and selection is encoded as a decrease in firing rate. For these projections, the following learning rule is used:

$$\tau_w \frac{dv_{ij}}{dt} = f_{DA} \left( DA(t) - B_{DA} \right) C_{ij} - \alpha_j(C_{ij})$$

with $f_{DA}$ and $\alpha_j$ defined in Equation 3 and with:

$$C_{ij} = \left( r_i - \bar{r}_{PRE} - \gamma_{PRE} \right) \left( \bar{r}_{POST} - r_j - \gamma_{POST} \right)$$

for the projections in the dorsomedial loop and

$$C_{ij} = \left( Tr_i - \overline{Tr}_{PRE} - \gamma_{PRE} \right) \left( \bar{r}_{POST} - r_j - \gamma_{POST} \right)$$

for the projections in the dorsolateral loop. Parameter values for the function $f_{DA}$ in these connections are presented in Table 2. All connections are initialized with positive random values drawn from a Normal distribution with mean 0.1 and standard deviation 0.02. The parameters for these projections are presented in Table 3.

Projections to the subthalamic nucleus follow the same rule as those to striatal D1 expressing cells (Equation 2 and $T_d = 1$) and projections from the subthalamic nucleus to the GPi follow Equation 7 and $T_d = -1$.

Two different dopaminergic cells project to each of the two loops. Each cell in the dorsomedial loop is associated with one of the two possible rewards. Each cell in the dorsolateral loop is associated with one of the two possible final states (east and west ends of the maze). These cells project back to all the nuclei of the corresponding loop. The activation is governed by the following equation:

$$\tau \frac{dm_j}{dt} + m_j = P(t) \left( R(t) + Q(t) \sum_{i@StrD1} w_{ij}r_j \right) + B$$

At resting conditions these cells are at a tonic level ($B = 0.1$) and have phasic changes only after executing an action. This is controlled through a function $P(t)$, which is 1 only after an action and 0 otherwise. On cells projecting to the dorsomedial striatum a phasic increase is only produced if reward is obtained. The size of this rise in firing rate is controlled by the function $R(t)$, which is $(1 - B)$ if a reward is obtained and 0 otherwise. On cells projecting to the dorsolateral striatum, a phasic change with $R(t) = (1 - B)$ is produced after every action if an environmental state associated with the cell is achieved, independent if reward was received or not. Following the reward prediction error hypothesis, the response of dopaminergic cells is reduced once the reward or reached state can be fully predicted based on striatal D1 cells. This is achieved in the model through inhibitory plastic connections from the striatum which grow with successful trials

### Table 2 Parameters of the plastic connections between the cortex and the basal ganglia

| Parameter     | Dorsomedial | Dorsolateral |
|---------------|-------------|--------------|
|               | Str. D1    | Str. D2     | STN | Str. D1 | Str. D2 | STN | IL |
| $T_d$         | 1          | -1          | 1   | 1       | -1      | 1   | - |
| $K_b$         | 1.0        | 1.0         | 1.0 | 1.2     | 1.0     | 0.6 | - |
| $K_d$         | 0.05       | 0.2         | 0.4 | 0.05    | 0.4     | 0.4 | - |
| $\gamma_{PRE}$ | 0.35       | 0.2         | 0.15 | 0.1     | 0.1     | 0.15 | 0.0 |
| $\gamma_{POST}$ | 0.0       | 0.05        | 0.0 | 0.0     | 0.0     | 0.0 | - |
| $\tau_w$      | 100        | 10          | 1,500 | 600     | 60      | 1,000 | 9,000 |
| $\tau_a$      | 2.0        | 2.0         | 1.0 | 15      | 1.0     | 1.0 | 10.0 |
| $m_{MAX}$     | 1.0        | 1.5         | 1.0 | 0.9     | 1.0     | 0.4 | 3.0 |
and can cancel the effect of \( R(t) \). Additionally, to fully recreate the reward prediction hypothesis, the dopamine signal has a strong drop below baseline when reward is predicted but not received. To produce this behaviour, \( Q(t) \) in the dorsomedial loop scales differently, so \( Q(t) = -1 \) for a rewarded trial and \( Q(t) = -10 \) for an unrewarded trial. For the dopaminergic cells projecting to the dorsolateral striatum, \( Q(t) \) is fixed to \(-1\).

Learning on the projections from the striatal cells to the dopaminergic cells will increase their inhibitory effect after every successful trial. The phasic increase in dopamine will be therefore reduced after each successful trial controlling learning in the BG. Plasticity is then governed by:

\[
\tau_w \frac{dw_{ij}}{dt} = g_{DA} \cdot (DA(t) - B_{DA}) \cdot (r_i - \bar{r}_{PRE})^+ \quad (11)
\]

where:

\[
g_{DA} = \begin{cases} 
1 & \text{if reward} \\
3 & \text{if no reward} 
\end{cases} \quad (12)
\]

for the dorsomedial cells and \( g_{DA} = 1 \) for the dorsolateral cells, \( w_{ij} \) is the weight between the striatal cell \( i \) and the dopaminergic cell \( j \), \( DA(t) \) is the dopamine level in the corresponding loop (sum of dopaminergic inputs to the striatal cell \( i \)), \( r_i \) is the rate of the presynaptic cell, \( \bar{r}_{PRE} \) is the mean of the presynaptic layer and \( \tau_w \) is a time constant (3,000 for the dorsomedial loop and 12,000 for the dorsolateral loop).

Plasticity in the projections between the stimulus cells and the IL layer is not modulated by dopamine and therefore follows a Hebbian learning rule:

\[
\tau_w \frac{dw_{ij}}{dt} = C_{ij} - \alpha_j \left( (r_j - \bar{r}_{POST})^+ \right)^2 w_{ij} = C_i (r_i - r_{PRE} - \gamma_{PRE}) + (r_j - \bar{r}_{POST} - \gamma_{POST}) \quad (13)
\]

Learning in the IL is much slower than in the striatum (\( \tau_w = 9,000 \) in the IL and \( \tau_w = 100 \) in the striatum) as these connections are not modulated by a dopamine signal. Therefore, in order for the IL to acquire a pattern, it is necessary that the basal ganglia make the same selection multiple times.

During early trials, all cells in the IL activate at a similar level providing the thalamus with only unspecific activity. However, this unspecific activation enables the BG to bias a particular selection by its inhibitory influence on the thalamus. In the models with no shortcut, the baseline and the intrinsic noise of thalamic cells in the second loop were increased to compensate for the reduced excitatory input to the thalamus. Later, after enough training of the model with shortcut, the IL will bias selection without incorporating the goal signal.

The final selection is done stochastically according to a probability distribution computed using the activity of the cortical cells in the dorsolateral loop:

\[
P(a_i) = \frac{r_i}{\sum_i r_i} \quad (14)
\]

where \( P(a_i) \) is the probability of selecting the action associated with the cortical cell \( i \) and \( r_i \) is the firing rate of the cortical cell \( i \).

All differential equations are numerically solved using the Euler method with a time step of 1 ms using the neural simulator ANNarchy version 4.6 (Vitay, Dinkelbach, & Hamker, 2015). The value of all fixed connections is presented in Table 4.

### 2.3 Quantification and statistical analysis

We have measured changes in the connectivity by using the Bray–Curtis measure of dissimilarity between weight matrices as given by the following equation:

### Table 3 Parameters for the connections from the striatum to the GPi and GPe

| Parameter | Dorsomedial | Dorsolateral |
|-----------|-------------|-------------|
| \( T_d \) | 1 | -1 |
| \( K_b \) | 2.4 | 2.4 |
| \( K_j \) | 1.8 | 0.2 |
| \( \gamma_{PRE} \) | 0.15 | 0.2 |
| \( \gamma_{POST} \) | 0.3 | 0.0 |
| \( \tau_w \) | 550 | 600 |
| \( \tau_a \) | 20 | 20 |
| \( m_{MAX} \) | 1.2 | 1.5 |

### Table 4 Fixed connection values

| Presynaptic | Postsynaptic | Value dorsomedial | Value dorsolateral |
|-------------|--------------|-------------------|-------------------|
| GPe         | GPi          | 1.0               | 0.1               |
| GPi         | Thalamus     | 2.0               | 1.0               |
| Cortex      | Str. Feedback| 1.2               | 1.0               |
| Thalamus    | IL           | 0.05              | 0.05              |
| Str. Feedback | GPi     | 1.1               | 0.1               |
| Str. Feedback | GPe   | 0.3               | 0.15              |
| STN         | STN          | 0.3               | 0.3               |
| Str. D1     | Str. D1      | 1.0               | 1.0               |
| Str. D2     | Str. D2      | 0.3               | 0.3               |
where $u$ and $v$ are the two matrices whose dissimilarity is being measured. This value is 0 if both matrices are the same and 1 if they are completely different.

We have also computed the selectivity of cells to either an action or a goal by using the following equation:

$$S(x) = \frac{\mu_x}{\mu_x + \mu_y}$$

where $x$ is either an action or a goal, $\mu_x$ is the mean of the activity of the cell during trials in which the corresponding action was executed or the corresponding goal obtained and $\mu_y$ is the mean activity on the trials where the remaining action was executed or the remaining goal was obtained. The mean activity was first computed over time for each trial and then over trials.

### 3 | RESULTS

#### 3.1 | Effects of devaluation

We tested the neuro-computational model on the devaluation task used by Smith and Graybiel (2013) to study the shift from goal-directed to habitual behaviour. In this task, a gate opens and rats run in a T-maze as they have to select one end arm to receive reward. On each trial, either chocolate milk or a sucrose solution is available. Each type of reward is associated with one end arm and an auditory cue at the beginning of the trial indicates which reward is available.

Rats were exposed to a devaluation protocol early or late during training. Devaluation has been obtained by pairing one type of reward with a nauseogenic dose of lithium chloride in the home cage. If devaluation was done early during training, the performance dropped to 50%, but only in those trials in which the devalued reward is present. If the devaluation was done later, the rat did not change its behaviour and continued to select the location of the devalued reward.

In order to test the model on this task, a particular goal signal (R1 or R2), which encodes the desire for one particular reward (chocolate milk or sucrose), was provided to the dorsomedial network. The selection process for the desired reward given the tone is assumed to be solved by the limbic network. Each goal signal has an excitatory effect on half of the cells in the dorsomedial striatum and a feedforward inhibitory effect on all others, the latter representing the result of direct cortical projections to fast-spiking, parvalbumin-containing striatal interneurons (Kita, Kosaka, & Heizmann, 1990; Moyer, Halterman, Finkel, & Wolf, 2014; Parthasarathy & Graybiel, 1997).

The sensory input is encoded by 7 cells: 4 encoding the two possible auditory cues (2 per cue) and three encoding the opening of the gate. Learnable connections were established from these 7 neurons to all cells in the dorsomedial striatum and to the IL. Figure 3 illustrates how the devaluation task is implemented in the model. At the beginning of every trial the baseline of the active sensory input cells was set to 0.5 and of the active goal signals to 1.0 for 600 ms.

The thalamic and pallidal cells of the dorsomedial loop represent the two locations (state) of the environment where reward could be found, which correspond to the end of the east and west arms of the maze. The interaction between the dorsomedial and dorsolateral loops is achieved through a pair of cortical cells that encode the same two locations but that will receive an additional excitation when the east or west location has been reached, after the decision to turn. Plastic connections were included between these cells and all neurons in the dorsolateral striatum, allowing for learning the appropriate action which leads to the goal state.

The two cortical cells in the dorsolateral loop represent a turn direction (left and right), and to allow some variability, their activity is transformed into a probability distribution using a soft-max rule. The final decision is determined stochastically according to this distribution. Devaluation was simulated by cancelling the activation of the corresponding goal signal. This can be interpreted as the lack of interest of the animal in obtaining the devalued reward.

We trained two versions of our model, a full version and one in which we removed the shortcut, to execute the correct action (or turn) given a goal signal and sensory input. Initially, the models selected actions randomly, reached a performance of 80% after 30 trials and further saturated at a level of 90% (Figure 4, top row), similar to what was observed by Smith and Graybiel (2013) for rats. After training, cells in the dorsomedial striatum were selective for combinations of auditory cue and goal (Figure 5). These cells became linked to the particular GPi neuron that disinhibits those neurons in the thalamus and cortex that are associated with the spatial location of an end arm where the reward is usually presented. This selection of a goal location guides the appropriate response selection in the dorsolateral loop.

When testing the models on the devaluation protocol, we trained the models to criterion (until each reaches on average a performance of at least 90% within the last 10 trials). Then, the devaluation condition comprised 10 more trials with each goal/tone combination. On those associated with the devalued reward, no goal signal was activated. Devaluation reduced the performance to about 50% on trials associated with the devalued reward, again independent of the shortcut. However, if devaluation was performed in an overtrained condition (80 trials after the 90% condition was met), devaluation was not effective in the models with shortcut (see Figure 4). These results compare well to those of rats (Smith & Graybiel, 2013). In the models without a shortcut, devaluation is effective, what
is different from the observations of Smith and Graybiel (2013). This suggests that the concept of a shortcut can serve as a reasonable explanation for the development of habitual responses.

This change in response becomes clear when we analyse the activity pattern of the dorsomedial striatal cells (Figure 6a). Under normal conditions, striatal cells tuned to the corresponding goal show a high activation, while others kept a low level of activity. After devaluation, the lack of a goal signal produced a pattern in which cells that are tuned to the paired reward had no advantage anymore. As a result, the BG does not produce a strong objective signal for a preferred state. This unbiased input to the dorsolateral striatum produces an almost random selection, reducing the performance to about 50%. After overtraining (additional 80 trials of training), however, the IL became sufficiently selective and biased the dorsolateral network to overcome the devaluation. This behaviour is evident as the striatum quickly develops a strong activation (Figure 6b), while the IL requires a large number of trials (due to slow learning). Similar activation patterns were recorded by Smith and Graybiel (2013) using chronic tetrodes implanted in both the striatum and the IL of rats.

### 3.2 Simulations of a place/response task

In a second set of experiments, we tested the model on a place/response learning task that was initially used by Packard and McGaugh (1996) to study the different roles of the hippocampus and the caudate nucleus during learning and decision-making. In this task, a rat is placed into the south arm of a cross maze with the north arm closed. Reward is given in one of the two remaining arms by placing food pellets at the end of it. During training, the animals need to learn which of the two arms is the rewarded one by exploring both and initiate the correct turn direction.

Rats were then injected either a lidocaine or a saline solution (sham) in either the caudate nucleus or the hippocampus at the end of two different stages of training. Lidocaine injections are assumed to inactivate neural tissue during a limited period of time (Tehovnik & Sommer, 1997). The north arm of the maze was opened and the south arm closed. Rats were placed in the north arm and the arm entered by the animals (east or west) was observed. If the rat entered the non-rewarded arm by choosing the same turn action learned during previous training, the rat was considered a response learner. If instead the rat entered the rewarded arm, realizing that it was in a different starting position, the rat was considered a place learner.

Rats that received a saline injection (sham) in either the hippocampus or the caudate were mainly place learners when the test trial was performed at day 8 but were mainly response learners when the test trial was performed at day 16. Rats with a lidocaine injection to the caudate were also mainly place learners when the test was performed at day 8 but remained so even when the test was performed at day 16. Half the rats with a lidocaine injection in the hippocampus were place learners and half were response learners if the test trial was performed at day 8; however, if the test trial was at day 16, most rats were response learners. This observation has been interpreted on
basis of two separate learning systems, an allocentric-based hippocampal and an egocentric striatal system.

To test the model with this task, we used a similar configuration as the one used for the previous devaluation protocol, but included only a single goal as only one type of reward is available on every trial. The output neurons of the dorsomedial loop encoded again the two states (the end of each arm). To model the influence of the hippocampus, we included an additional signal to the thalamus of the dorsolateral loop during the test trial in order to bias the turn direction towards the correct place. The signal may represent the results of a computational process which is likely performed by multiple cortical areas together with the hippocampus that detects uncertainty in the environmental conditions and applies place knowledge to propose responses (Durstewitz, Vittoz, Floresco, & Seamans, 2010; Murphy, Mondragon, & Murphy, 2008; Stefani & Moghaddam, 2006).

As the lidocaine injection to the caudate may have affected the dorsolateral striatum but potentially also the dorsomedial striatum, we tested the effects on either, the dorsomedial and the dorsolateral loop in our model. To simulate a lidocaine injection to the caudate, we set the rate of all striatal neurons of the respective loop to zero, independent of the input received by these cells. To simulate a lidocaine injection to the hippocampus, we found that a reduction of the strength of the additional input integrated during the test trial by 7% accounts well for the data, as still half of the rats are place learners with a hippocampus...
lesion. This implies incomplete lesions or that the hippocampus is not the only source of the allocentric bias received by the BG.

When trained on the task the model reaches a performance of 90% after 42 trials (Figure 7a). As in the experiments of Packard and McGaugh (1996), if the test trial is performed early, here at trial 43, most models act as place learners; however, if the test is performed after longer periods of learning, here at trial 97, most models act as response learners due to the learned bias of the shortcut. When we suppressed the activity in the striatum of either of the two loops (putative effects to lesioning the caudate nucleus), models rarely become response learners, as was observed by Packard and McGaugh (1996) with rats. Alternatively, if we reduce the hippocampal signal, half the models act as place learners and half as response learners when lesioned early (trial 43), but if the lesion is performed late (trial 97) most models act as response learners (Figure 7b). This also compares well with the results of Packard and McGaugh (1996).

While normal models become response learners after enough training of the shortcut, models with a lesion in the striatum remain place learners, as dorsomedial or dorsolateral lesions impair the suppression of the hippocampal signal. Under normal conditions, during the training period, the dorsolateral loop learns to inhibit the turn direction that was unrewarded by the hyperdirect pathway and to select the turn direction that was rewarded (Figure 8a). However, this effect is only produced if the dorsomedial loop has previously selected an expected state and the dorsolateral striatal cells are active. When the test trial is performed early, this hippocampal excitatory signal is strong enough to surpass the inhibition from the hyperdirect pathway and the hippocampal-biased action is selected more often (making model place learners). If the test is performed late, two effects add up and the previously rewarded action is selected more often (making models response learners). First, the IL mediated habitual response develops and second, the inhibition of alternative actions from the hyperdirect pathway impairs the activation of hippocampal-biased action. Models with a lesion in the dorsomedial striatum show a lower inhibition in the dorsolateral basal ganglia (see Figure 8c), allowing the hippocampal-biased action to gain influence. Models with a lesion in the dorsolateral striatum do not disinhibit the thalamus, therefore rendering alternative actions, than the hippocampal-biased one, unlikely (see Figure 8c).

3.3 | Computational benefits

Until now we primarily addressed the reproduction of experimental data by the model and its underlying mechanisms. We now shift to outstanding questions that are more computational in nature, ask for the potential benefit of such a hierarchical organization of decision-making and provide evidence for an improved transfer of knowledge across tasks.

Keeping with the T-maze setup, we ask whether the knowledge obtained when finding one reward can be transferred to finding another reward, even if it is at a different location than the previous reward. We initially trained the model on the task of Packard and McGaugh (1996) where one type of reward was associated with a single arm of a T-maze. From trial 30 onwards, we removed the reward and incorporated a second type of reward, located at the previously unrewarded arm (Figure 9). The switch was informed to the network by a change in the goal signal reaching the dorsomedial loop, which now indicated the new reward, but of course not its location. We expected the model to quickly learn the second task, if our implementation of the basal ganglia loops can reuse information obtained during the unrewarded trials of the first task. Thus, only if it can acquire generic world knowledge during exploration, it can benefit under this new condition from previous experience. To compare with naive learning conditions, we generated another model where we reset the weights on the cortico-striatal connections of the dorsolateral loop, removing the information gathered in this projection during the first task.

While the model can learn the second task in either condition, models that have been trained on the previous task require much less trials (Figure 9). Models following our hierarchical organization of decision-making are able to transfer the information from the first task to the next as the second loop already learned how to reach a particular state,
for example east arm during the unrewarded trials. The naive models, however, need to learn both the outcome of actions and the position of the new reward.

A particular assumption of our model is that the dopamine signal is different for the two loops, informing of the errors at the appropriate abstraction level. However, previous computational models with multiple loops have shown that it is possible to learn a task using a single dopamine signal (Collins & Frank, 2013; Frank & Badre, 2012; Schroll, Vitay, & Hamker, 2012). We therefore wondered if our new concept could provide a computational benefit and compared the full model with a version of the model in which both loops receive just the reward prediction error dopamine signal. Additionally, we tested the effect of the shortcut in any of the two versions.

We initially tested the models in the same task, in which only one reward is available in a single arm of a T-maze. The model with different prediction errors for
each loop clearly outperforms the model with an identical (reward) prediction error in both loops (Figure 10). When the shortcut was removed, the model with an identical prediction error could not even solve this simple task. The problem with models composed of multiple loops learning by an identical prediction error signal is that not individual decisions but only the final outcome is observable. In the context of our task, reward R1 has to be associated with a decision to go to the East or West arm. However, if no error signal about this decision is considered for learning and only the outcome after the turn left or right action is used, neither the decision to go East nor that to go West is dominant and both get equally often rewarded or punished, constantly (but randomly) switching between LTP and LTD and this impairing the convergence of weights in the medial BG. A shortcut helps, as it will bypass this first loop and introduces a bias in the final decision. So we hypothesized that when we design our task being more complex, it could surprisingly be learned.

We therefore further tested the models with an identical prediction error in both loops in a probabilistic rewarded task, in which two rewards were available. On 75% of the trials one reward was available in one arm, while on the remaining 25% a different reward was available in the other arm (see Figure 10). The available reward was selected randomly and informed to the model only by activating the corresponding goal signal reaching the dorsomedial loop. The models should therefore learn to first select the appropriate arm depending on the available reward in each trial and then to choose the proper action to reach it. Now also the model without a shortcut can solve the task. This demonstrates that models composed of multiple loops learning by an identical prediction error signal in all loops do have a fundamental learning problem, but depending on the task, it may not become apparent.

3.4 | Prediction: different activity patterns on each loop after relearning

So far, we have evaluated the model with experimental data and tested its computational properties. We now turn to two central model predictions to reveal the full potential of the model. First, our approach suggests that the dorsolateral loop can learn appropriate actions under multiple contexts, each with different action–outcome associations. Second, our hierarchical approach suggests that the dorsomedial network selects a state which allows to obtain a desired goal (reward) and predicts therefore that neurons are selective for states and not necessarily for actions. Similarly, the dorsolateral network learns to map states to responses and predicts therefore that dorsolateral striatal cells become selective to the executed action and not necessarily to the reached state. In the task of Smith and Graybiel (2013), however, there is only one context and each action is always associated with the same state (and reward); hence, action and state selectivity are indistinguishable. For these
reasons, we have implemented a similar task (based on the task of Smith & Graybiel, 2013) in a flexible environment which allows changes in action–state and stimulus–reward location mappings.

Prior to the trial, we selected randomly one of two possible environments, each with a T-maze but with enough sensory clues to easily differentiate between them. Importantly, the environments differ by the outcomes of the action taken (see Figure 11a). After 30 stable trials, we swap the action–outcome rule in the environment (action switch) or the position of the rewards (goal switch). Although this setup is physically impossible as turning to the same side will always lead to the same arm (given the same starting point), it can be implemented by virtual reality technologies for freely moving rats (Thurley & Ayaz, 2017). New technologies, as the apparatus used by Kaupert et al. (2017), include a spherical treadmill controlled through a closed loop which allows to adapt the scenario according to the animals’ decisions.

While either an action switch or a goal switch requires the model to discover the new rule, the model predicts that an action switch will produce mainly changes in the dorsolateral loop and a goal switch will produce mainly changes in the dorsomedial loop. In simulations with a goal switch, each action still leads to the same state, and therefore, no change is produced in the dopamine signal reaching the dorsolateral loop. However, the dopamine signal reaching the dorsomedial loop decreases (expected to receive reward but was not obtained) and forces the combinations of goal and stimuli represented in the striatum to be linked to a different intended state in the GPi. In simulations with an action switch, the state reached after executing an action is unpredicted, and therefore, the dopamine signal of the dorsolateral loop increases. This enhances plasticity between the active cortical cell, which is the one encoding the new reached state, and the striatal cells encoding the selected action (see Figure 11b).

Figure 12 shows that the model can learn the initial rules in the two different mazes (maze information is given
as input to the dorsolateral loop) and successfully adapt after both an action and a goal switch. Although the performance in both cases was similar, the changes produced in the model were different. As predicted, on simulations with an action switch the main changes occur in the dorsolateral loop, particularly in connections between the cortex and the dorsolateral striatum. On simulations with a goal switch, the main changes appear in the dorsomedial loop, especially in connections between the dorsomedial striatum and the GPi.

Our predictions could be tested by computing the selectivity of dorsomedial and dorsolateral striatal cells to both the achieved reward and the performed action. After any of the two types of switches, medial striatal cells which were initially active mainly on trials in which one of the two rewards was available will stay active in the trials associated with the same reward, even though the performed action has changed. These cells will therefore maintain their selectivity to the achieved reward but change their selectivity to actions (Figure 13). Lateral striatal cells, which were initially active in trials in which one of the two actions was taken, will remain active on trials in which the same action is selected, even though the state and reward achieved are different. These cells will therefore maintain their selectivity to the performed action but change their selectivity to the achieved reward (Figure 13).

4 | DISCUSSION

We have introduced four key concepts relevant for the organization of multiple cortico-basal ganglia-thalamocortical loops and verified them by means of a neurocomputational model in the context of habit learning. Our hierarchical approach suggests that a combination of a desired goal and environmental information triggers the selection of an objective in the dorsomedial network (caudate) which will then prompt the selection of an appropriate action in the dorsolateral network (putamen). Habits emerge as a consequence of slow plasticity in a shortcut that bypasses the dorsolateral network and thus biases a selection purely based on stimulus information.

Although habitual and goal-directed behaviour have been previously associated with parallel circuits implementing a model-free and a model-based reinforcement learning algorithm (Daw, Niv, & Dayan, 2005), no clear evidence for this strict dissociation in the brain exists (Miller, Ludvig, Pezzulo, & Shenhav, 2018; Pennartz et al., 2011). Evidence indicates that brain areas that were initially associated with model-based learning also support model-free learning (Doll, Simon, & Daw, 2012). This suggests that better neuro-computational models are required to illuminate the computations in the brain constrained by anatomy and physiology. We believe our approach is a step forward in this direction.

4.1 | Impact of reward devaluation and lesions on behaviour and neural activity

Previous theories, based on two parallel learning systems, suggest that devaluation has an effect only if it is performed before control has been given to the habitual system (implemented in the dorsolateral BG) by the goal-directed system (implemented in the dorsomedial BG). However, the exact mechanisms by which control is switched are unclear. Our model does not require an explicit switch of control. The inefficiency of reward devaluation after long periods of learning occurs, as the shortcut bypasses the high-level loop and thus
renders the model ineffective regarding changes at the level of high-level motivational changes. A devaluation procedure has therefore an effect only, if it occurs before the shortcut is sufficiently trained. Our model replicated core behavioural data and also matched neural recordings in the infralimbic (IL) cortex of Smith and Graybiel (2013).

Our hierarchical approach also provides a novel explanation of the effects of caudate lesions in the task of Packard and McGaugh (1996). Although a lesion on either loop reproduces the effects reported by Packard and McGaugh (1996), the mechanism on both cases are different. A lesion to the higher level of the hierarchy had only an indirect effect on the lower levels. Dorsomedial lesions produce anomalous objective signals to the dorsolateral network and impede its normal operation. Under normal conditions the dorsolateral BG is capable of inhibiting alternative actions but after a lesion of the dorsomedial striatum, the dorsolateral loop lost this bias. Models with a lesion of the dorsolateral striatum do not remove the tonic inhibition in the thalamus, and therefore, any action different than the hippocampal-biased one becomes improbable. Our results suggest that cells from any of the two loops could have been encompassed by the injections performed by Packard and McGaugh (1996).

Lesion experiments have been classically performed to investigate the computational functions of different striatal regions. Evidence for the parallel learning system model builds upon observations that lesions to the dorsomedial striatum made behaviour habitual even after a short training period (Yin, Ostlund, Knowlton, & Balleine, 2005) and that lesions to the dorsolateral striatum transform habitual responses into goal-directed behaviour (Yin, Knowlton, & Balleine, 2004). Our model is consistent with these findings. As habits are represented as cortico-thalamic shortcuts, they are still under control of the basal ganglia: the shortcut increases the activity of particular cells in the thalamus which will activate a closed thalamus–striatum–GPi–thalamus loop that will further enhance any external activation of the thalamus due to the disinhibition by the basal ganglia. A striatal lesion of this loop will abolish any disinhibition so that the thalamus will remain under inhibitory control reducing the effect of the shortcut. Thus, according to our model, a lesion to the dorsolateral striatum predicts a reduced effect of the shortcut.

**FIGURE 10** Comparison between the performance of models with a single dopamine signal and different dopamine signals. When a single reward is available (left) the models with two dopamine signals learn faster than those with just one signal. Without shortcut, the model with a single reward prediction error signal for both loops could not even learn the task. In a more complex task, in which two rewards are available (right), one on 75% of the trials and one on the remaining 25%, the difference between models is less pronounced [Colour figure can be viewed at wileyonlinelibrary.com]
Lesion to the dorsomedial striatum will impede the transfer of an objective to the dorsolateral loop which will then have no influence on the thalamus of the dorsolateral loop. The action selected by the shortcut will therefore have no competing, alternative actions, which makes behaviour more dependent on the shortcut.

While we addressed the transition to habitual behaviour mainly at the systems level of brain organization, our account is not inconsistent with recent observations at the macro- and micro-level of neural organization, although an exact comparison would require a more detailed model of the striatum. O’Hare et al. (2016) showed that habitual behaviour correlated with an increased output of the dorsolateral striatum and a shift in the relative timing of the indirect and direct pathway striatal cells. In mice expressing habits, the direct pathway cells tended to fire first. The spike probability of striatal cells however did not correlate with habitual behaviour. Although in our simplified tasks an activation of the indirect pathway is hardly required to suppress alternative actions in our model, the overall output increase could be produced by an increased influence of the shortcut in mice expressing habits, which through thalamic feedback would more strongly affect the firing of striatal cells, shaping their activation pattern. Given a faster signal transmission of the shortcut this may also explain the faster response of direct pathway striatal cells. At the microcircuit level, O’Hare et al. (2017) showed that fast-spiking interneurons are more excitable in mice expressing habits and that an acute chemogenetic inhibition of fast-spiking interneurons prevents the expression of habitual behaviour but not lever-pressing per se. Further, their in vivo recordings showed that, although fast-spiking interneurons exert a strong inhibitory influence over dorsolateral striatum output, they promote activity in a small set of spiny projection neurons. In our model, we represent the effect of fast-spiking interneurons as a broad range feedforward inhibition of those striatal cells which represents goals differing from the currently desired one. This produces a similar effect on striatal neurons as observed by O’Hare et al. (2017) in their recordings, where most cells are inhibited and a small group

**FIGURE 11** (a) Diagram showing the proposed new task used to test the full potential of the model. The task uses two distinct T-mazes which can be differentiated by their visual features (in the diagram with different initial arm colours). Each maze has the same reward on each of the two possible and recognizable end states (in the diagram with solid and dashed lines). To reach each state, however, a different action is required on each maze. At the beginning of each trial, one of the two mazes is selected randomly and the rat is placed on it. After 30 trials, either the state reached after executing each turn is changed (action switch), or the position of the two rewards changed (goal switch). (b) If the consequence of actions are switched and the animal reaches the solid state instead of the dashed state after turning to the left, a dopamine neuron of the dorsolateral loop which encodes the new state prediction error is activated. As this cell has never been active on trials in which the animal has turned left in the current maze, this dopaminergic cell receives only low inhibition from the active striatal neurons. This dopamine peak enhances plasticity between the active state in the cortex (solid line) and the striatal cells associated with the performed turn direction, forcing a change in the state-action associations learned in the dorsolateral loop [Colour figure can be viewed at wileyonlinelibrary.com]
shows enhanced activity. Further, any disruption of the feedforward inhibition would increase the general level of activity of the striatum, increasing the influence of goal-directed behaviour expressed through the BG loops and reducing the influence of habitual behaviour expressed through the shortcut. Additionally, an increased inhibitory activity during habitual behaviour, as observed by O’Hare et al. (2017), could decrease the influence of the cortico-striatal pathway and indirectly enhance the one of shortcuts.

4.2 Limitations

Despite the consistency of our model with central observations, the model is kept simple and thus does not replicate some observations. For example, the neural measurements of Smith and Graybiel (2013) show that both the dorsolateral striatum and the IL have high activity only at the beginning and end of each run, referred to as bracketing activity. Our numerical experiments do not show the same pattern mainly because the task simulated is a simplification of the protocol performed by Smith and Graybiel (2013) with rats. In our procedure, reward is delivered immediately after the decision to turn. In the rat experiment instead, animals must finish running before obtaining the food. In the animal experiment, the intended state is therefore not reached immediately but after a few seconds. The two peaks may correspond to the initial decision and to the feedback that indicates that a final state has been reached (the two inputs to the cortical layer between the two loops). In our simulations, both peaks occur together.

Data have also shown that the dorsomedial striatum contributes preferentially to the early learning period while the dorsolateral striatum to the late learning period (Miyachi, Hikosaka, & Lu, 2002; Yin et al., 2009). We hypothesize that a more realistic signalling from the limbic network may explain this. The observed activation patterns could be explained by a reduction of the goal signal late during training, once the shortcut has been fully trained. However, in our simulation, the goal signal is constant during the complete simulation. A reduction could be useful to reduce energy consumption once a behaviour has become habitual. This

FIGURE 12 Performance and changes in the model after a switch in the environmental conditions. The top row shows the performance of the model on simulations with a switch in the consequence of actions and on simulations with a switch in the position of goals (or rewards). Both switches were performed at trial 30 without informing the model. In both cases, the model learns the initial associations in less than 30 trials; then, performance drops to 0 as the conditions have changed, but the model continues selecting the previously correct action. After trial 80, the models have learned the new associations, reaching again high performance. The bottom row shows the Bray–Curtis dissimilarity between the weight matrices before the change in the environmental conditions and at the end of the simulation. When the consequence of actions is changed, the dissimilarity is higher for the projections from the cortex to the lateral striatum than from the projection from the cortex to the medial striatum and from the striatum to the pallidum. When the position of the goals is changed, the connections from the striatum to the internal pallidum of the dorsomedial loop are higher than all others. [Colour figure can be viewed at wileyonlinelibrary.com]
would mainly affect the dorsomedial striatum, which directly receives the goal signal as an input, and not the dorsolateral striatum, which will be activated anyway through thalamic feedback by the shortcut.

4.3 Organization of cortico-striatal loops in habit learning

Our model differs from most previous theories of habit learning through the assumption of overlapping cortico-striatal projections for communication between loops. Previous theories proposed the use of the striato-nigro-striatal network (Manella, Mirolli, & Baldassarre, 2016; Yin, 2014; Yin & Knowlton, 2006) which only allow to transfer the results of the striatal computations. Our approach allows the loops to transfer the selection obtained by the output nuclei of the basal ganglia. Another option which could provide a similar functionality to overlapping cortico-striatal projections is the transfer of information through the cortico-thalamic-striatal pathway (McFarland & Haber, 2002).

Manella et al. (2016) developed a model for value learning dependent on the internal state of an animal and applied it to devaluation experiments in instrumental learning tasks. Although the model is composed of three loops (one ventral and two dorsal), learning is simplified as cortico-cortical connections that connect the loops are not adaptive, thus having fixed goal–action associations. The role of the loops is then restricted to implement competition between possible responses. Due to the narrow learning capabilities in the cortex–basal ganglia loops, decision-making is limited to the selection of goals in the ventral loop which includes the basolateral amygdala. While Manella et al. (2016) address devaluation experiments, they do not attempt to explain the emergence of habitual behaviour with overtraining. However, as the goal selection is simplified in our approach, we believe both models could be considered complementary in their focus of addressing goal-directed behaviour.

In summary, rather than parallel learning systems, we propose that habitual behaviours occur by means of the development of cortico-thalamo-cortical shortcuts. These shortcuts are trained and supervised by the basal ganglia.

4.4 Learning mechanisms in multiple cortex–basal ganglia loops

Learning in multiple cortex–basal ganglia loops leads to a spatial credit assignment problem. During training, a double error may occur, such that a loop may make a wrong selection that could lead to the initial goal even if an incorrect intermediate objective was selected by a previous loop or could lead to no reward even if the correct intermediate objective was selected previously. Solely learning from the selected objective could therefore lead to incorrect connectivity patterns, which even occurs on simple tasks (Figure 10). This fundamental limitation, which is inherent to all models that learn in two or more loops with an identical prediction error.
signal, can be hidden by a biased task design as illustrated in our results or biased model parameters or different speeds of learning in each loop. Thus, learning in multiple cortex–basal ganglia loops imposes a serious problem that has not been solved so far.

We here proposed that separate prediction error signals can be a solution. In our novel approach, the cortical cells integrate both environmental information and the selection made by the BG, allowing the model to solve this credit assignment problem by learning from what was achieved given the available information in each cortex–basal ganglia loop. These distinct procedures allow the network to quickly discover the environment. The dorsolateral loop learns not only on trials in which reward was obtained nor only on trials in which the intermediate objective selected by the first loop was reached, learning occurs on every trial, increasing the strength of the association between the achieved intermediate objective and the executed action. This mapping can be exploited later by the dorsomedial loop if under different conditions it selects a particular objective which has already been mapped to an action during unrewarded trials. Such a capacity could be especially useful in environments in which the reward conditions change or when multiple rewards are available. It has been widely shown that mice that have been exposed to more complex environments perform better in several maze tasks (Lewis, 2004) an effect that could be explained by animals developing an objective–action mapping in the dorsolateral loop independent of the particular task. Although some explanations of the increase in performance point mainly to an increase in the number of neurons in the hippocampus (Garthe, Roeder, & Kempermann, 2016; Kempermann, Kuhn, & Gage, 1997), experiments have shown that animals exposed to enriched environments have higher metabolic activity and dendritic spine densities in the motor loop (Turner, Lewis, & King, 2003; Turner, Yang, & Lewis, 2002).

Finally, our model proposes the existence of shortcuts which are slowly trained by the basal ganglia. In a previous study, we have already shown that combining a slow and a fast learner could lead to an increase in the generalization capabilities of a single loop basal ganglia model (Villagrasa et al., 2018). An additional benefit of the shortcut structure we have proposed is that any automatic behaviour could be cancelled if the activation of the shortcut is balanced through top down control after the detection of a conflict. Further, Smith and Graybiel (2013) showed that if the rat stops its habitual behaviour, the IL returns to its initial activity pattern. This agrees with the conflict monitoring hypothesis (Botvinick, Cohen, & Carter, 2004; Isoda & Hikosaka, 2007), which proposes that the brain first detects the occurrence of conflict and then cancels the execution of automatic behaviour. In addition, fMRI data have shown that during a task switching paradigm, the activity of the pre-supplementary motor loop increased in trials in which the current task changed and a different stimulus–response mapping was required (Korb et al., 2017). This could correspond to a monitoring process performed by the BG which determines when shortcuts should be overruled. Shortcuts may not only exist between the dorsomedial and dorsolateral loop but between various loops which may explain the emergence of habitual control of goal selection (Cushman & Morris, 2015).

4.5 Model predictions for habit learning

As a way to test a core prediction of our hierarchical approach we propose to add a relearning protocol to the original task of Smith and Graybiel (2013). Our model predicts that after either a switch in the consequence of actions or in the position of the rewards, dorsomedial striatal cells will maintain their selectivity to goals but change their selectivity to actions and dorsolateral striatal cells will maintain their selectivity to actions but change their selectivity to goals. Rats could be initially trained in a similar fashion as in the original task and then the position of the rewards or the consequence of actions switched. Changes in the activation of dorsomedial and dorsolateral striatal neurons should be measured and compared.

The hierarchical structure further predicts that the selection of an objective in one loop prompts the activation of the following one. Such a spiral pattern of activation should be visible if measurements were performed in multiple striatal and PFC locations simultaneously. During initial learning, we expect to see an ordered sequence of activation moving from one loop to the next. However, late during training shortcuts may become active which may suppress this pattern. Granger causality (Dhamala, Rangarajan, & Ding, 2008) could also be measured between different loops. We anticipate a strong causality between adjacent loops.

4.6 Implications for and comparison to computational models of decision-making beyond habit learning

Although we have focused on habit learning, our proposed concepts of multiple cortex–basal ganglia loops relate to some previous computational work outside the domain of habit learning. While a variety of neuro-computational basal ganglia models of a single loop have been proposed that slightly vary in one or the other way (for reviews, see Frank (2011); Helie, Chakravarthy, and Moustafa (2013); Schroll and Hamker (2013)) just a few attempts have been made to provide computational evidence for the organization of multiple loops.
Our model proposes overlapping cortico-striatal projections for communication between loops, similar to Collins and Frank (2013). They designed a model composed of two cortex–basal ganglia loops to explain the learning of task sets. A high-level prefrontal loop learns to map a stimulus dimension to a task set (while the particular stimulus dimension, for example colour, that determines the task set has been predefined) so that a second parietal cortex loop combines the selected task set with additional sensory information, for example shape, to select an action. The combination of task set and stimulus occurs via a hard-coded prefrontal to parietal connection and via a plastic prefrontal to striatal connection. The latter is conceptually very similar to the overlapping cortico-striatal projection we use in our model.

Inspired by the organization of the prefrontal cortex and the ability of humans to learn complex (hierarchical) rules, Frank and Badre (2012) proposed a nested cortex–basal ganglia architecture for working memory gating and response selection applied to learning a hierarchical rule set, where the presence of one stimulus determines which dimension of another stimulus has to be evaluated to provide the response. The higher-level loop gates one stimulus in a memory slot, so that it provides the context to output gate a particular dimension (selective routing) to another cortical layer where the routed stimulus feature is used for response selection. They did not address the development of habitual behaviours.

In the model of Frank and Badre (2012) but also the one of Collins and Frank (2013), both loops learn with the same dopamine signal that represents a reward prediction error. In Frank and Badre (2012), however, the authors assume that active striatal cells boost the dopaminergic signal reaching them, making the dopamine level neuron specific. In comparison, we showed that learning with a single reward prediction error common across all levels of the hierarchy will severely impair learning as individual decisions are not observable as an error at the final outcome cannot be traced back to its origin.

Topalidou, Kase, Boraud, and Rougier (2018) implemented a model composed of a motor and cognitive loop whereby both interact with each other via cortico-cortical interactions. In between the motor and cognitive loop, an associative cortex and associative striatum allow a bidirectional information flow from the cognitive to motor and motor and cognitive loop. The model proposes a dual competition hypothesis according to which the final decision is derived jointly by the BG and by a cortical pathway and applied it to first selecting a target shape and then transferring it to a target location to determine a motor response. However, learning is only implemented in the cognitive loop (between cognitive cortex and striatum and between the cognitive and associative cortex). According to Topalidou et al. (2018), their model realizes no transfer from action–outcome to stimulus–response, but both systems collaborate, without a hierarchy, to jointly make a decision.

Schroll et al. (2012) designed a neuro-computational model of a cognitive and a motor cortico-basal ganglia-thalamo-cortical loop to learn a 1–2-AX task. The cognitive loop has the ability to selectively recruit and cancel working memory contents required for this task. The cortical parts of the cognitive loop inform the striatal part of motor loop about previous occurrences of relevant stimuli. Learning takes place simultaneously in both loops and the complex 1–2-AX task is learned by an incremental shaping procedure. A change of rules in the task can be accounted for by relearning.

### 4.7 Relation to reinforcement learning

Reinforcement learning (RL) has been rather developed from a machine learning perspective where an action brings the agent from a state $s$ to the next state $s_0$. The learning goal is to choose a sequence of actions at over time $t$ to maximize the resulting reward $R = r_t + r_{t+1} + \ldots$ (Sutton & Barto, 2018). However, RL has also been repeatedly related to learning behaviours in animals (Neftci & Averbeck, 2019). In model-free reinforcement learning, the agent learns values of states or state–action pairs, such as in Q-learning. Despite its powerful learning abilities, model-free RL has been criticized of being too inflexible with respect to multiple goals and changing environments. In model-based RL state transition, probabilities $p(s'|s,a)$ and reward probabilities $p(r|s,a)$ become stored as a world model to enable the agent to search in the model for a particular rewarding outcome $o$ or desired goal prior to selecting an action. This other extreme is very flexible, but requires a full world model which is problematic to obtain for large environments and contexts. Successor representations were proposed to store state transitions without the set of involved actions to evaluate the value of a state (Gershman, Moore, Todd, & Normal, 2012; Momennejad et al., 2012).

We have to make the comparison of RL with our model with some caution, as in its present implemented version, we just compute a single action for a given trial and typical RL tasks consider a sequence of actions. While typical model-free RL approaches allow to cash state–action values, but drop the particular outcome and replace it with a generic reward value, our approach keeps the outcome but also caches action values, so that our model remains goal-directed and sensitive to outcome devaluation. Compared to typical model-based approaches, which learn an explicit model and search in that model to find the action that leads to the expected outcome, we do not search as we cash the searches $p(\alpha|o, s)$. Although the model does not explicitly search for the desired outcome, it is still quite flexible, as it allows to transfer knowledge across the tasks (via the hierarchy of decisions) and can incorporate...
different contextual situations. Shortcuts take out some information from this world model and make it more model-free, so that it can become habitual. This may make our current implementation an interesting starting point to further develop this framework for more complicated tasks, which require sequences of actions.

The computations performed by the basal ganglia have also been compared to hierarchical reinforcement learning algorithms (Botvinick, 2012; Ito & Doya, 2011). In hierarchical reinforcement learning, a task is divided into sub-tasks and for each, a sequence of actions or sub-policy is learned (Barto & Mahadevan, 2003). Once such a sub-task is learned, this sub-task could be used as part of a new task. A critical open question is how the separation into sub-tasks could be realized in an automatic way. At present, a comparison of our model to hierarchical reinforcement learning must be also made with care as the different levels in our hierarchy do not necessarily learn sub-policies but an abstract representation of outcomes that can be accomplished by a single action. However, it appears feasible that our model could be extended to allow executing a sequence of actions.

5 | SUMMARY

We addressed four key concepts relevant for the organization of multiple cortico-basal ganglia-thalamo-cortical loops and verified them in the context of habit learning by neuro-computational methods. Inspired by anatomical observations of overlapping cortico-striatal projections (Groenewegen et al., 2017; Haber, 2016) the first key concept proposed a hierarchical organization of loops from goals, to cognitive, further to premotor loops and motor. However, this does not exclude potential other pathways between loops (Haber & Knutson, 2010; Joel & Weiner, 1994; McFarland & Haber, 2002). Functionally, our proposed organization has the advantage that the behaviour of the organism is executed under the context of a particular goal. As demonstrated in a maze task, a subdivision of the task into a sequence of decisions facilitates transfer learning from one task to another (Figure 9). The second key concept addressed which information is used for learning. As each loop defines an objective for the next loop, it could be considered for learning as well. However, we demonstrated that using the actual reached objective for learning allows to learn even in unrewarded trials (Figure 9), which has the substantial benefit of task independent learning of environmental–behavioural relationships. Shortcuts between loops have been introduced as third key concept and functionally been related to habit formation as demonstrated by simulations of two habit learning tasks (Packard & McGaugh, 1996; Smith & Graybiel, 2013) (Figures 4 and 7), providing an alternative explanation to the previously dominant theory of separate goal-directed and habitual cortex–basal ganglia loops. The concept of shortcuts also relates well to the putative role of the infralimbic (IL) cortex (Smith & Graybiel, 2013), as also supported by comparing IL neural recordings with recordings from neurons in our shortcut part of the model (Figure 6b). Finally, the forth key concept proposes that each loop shall compute its own prediction error signal for learning. This would suggest that the firing of DA neurons does not only indicate a reward prediction error, but dependent on their association to different loops, also other prediction errors. This key concept is probably the most speculative, but as such, may also be an important issue in the fields of cognitive control and basal ganglia learning to move on. We demonstrated that a model containing two cortex–basal ganglia loops has severe learning limitations when both loops learn by the same reward prediction error (Figure 10), but not, if each loop computes its own error. A further limitation occurs if learning is solely limited to obtain reward, as much useful learning can be already at the level of understanding relationships of actions and their consequences independent of any reward. Thus, there are good reasons for learning from different error signals in different cortex–basal ganglia loops. Anatomically, this is supported by the observation of a spiralling pattern in the striato-nigro-striatal projections (Haber, 2003). Further, it is acknowledged that the dopamine neurons do not only compute a reward prediction error but also code different properties (Engelhardt et al., 2019). We suggest that loop-specific prediction errors should be part of this code.

ACKNOWLEDGEMENTS

The authors would like to thank Julien Vitay and Lieneke Janssen for their helpful comments on previous versions of this manuscript. This work was supported by the European Social Fund at the Free State of Saxony (Grant ESF-100269974), by the German Research Foundation DFG HA2630/11-1 part of “Computational Connectomics” (SPP 2041) and by the BMBF project 01GQ1707 ”Multilevel neurocomputational models of basal ganglia dysfunction in Tourette syndrome.”

CONFLICT OF INTEREST

The authors declare no conflict of interests.

AUTHOR CONTRIBUTIONS

J.B. and F.H.H. designed the model. J.B. ran the simulations and analysed the data. J.B and F.H.H. wrote the manuscript.

DATA AVAILABILITY STATEMENT

The neuro-computational model has been programmed in Python using the neural simulator ANNarchy version 4.6 (Vitay et al., 2015). The software can be obtained on request.

ORCID

Fred H. Hamker https://orcid.org/0000-0001-9104-7143
REFERENCES

Alexander, G. E., DeLong, M. R., & Strick, L. P. (1986). Parallel organization of functionally segregated circuits linking basal ganglia and cortex. *Annual Reviews Neuroscience*, 9, 357–381.

Ashby, F. G., Turner, B. O., & Horvitz, J. C. (2010). Cortical and basal ganglia contributions to habit learning and automaticity. *Trends in Cognitive Sciences*, 14, 208–215.

Averbeck, B. B., Lehman, J., Jacobson, M., & Haber, S. N. (2014). Estimates of projection overlap and zones of convergence within frontal-striatal circuits. *The Journal of Neuroscience*, 34, 9497–9505.

Badre, D., & Nee, D. E. (2018). Frontal cortex and the hierarchical control of behavior. *Trends in Cognitive Sciences*, 22, 170–188.

Baladron, J., & Hamker, F. H. (2015). A spiking neural network based on the basal ganglia functional anatomy. *Neural Networks*, 24, 1–13.

Balleine, B. W., Killcross, A. S., & Dickinson, A. (2003). The effect of lesions of the basolateral amygdala on instrumental conditioning. *Journal of Neuroscience*, 23, 666–675.

Barto, A. G., & Mahadevan, S. (2003). Recent advances in hierarchical reinforcement learning. *Discrete Event Dynamic Systems*, 13, 41–77.

Botvinick, M. M. (2012). Hierarchical reinforcement learning and decision making. *Current Opinion in Neurobiology*, 22, 956–962.

Botvinick, M. M., Cohen, J. D., & Carter, C. S. (2004). Conflict monitoring and anterior cingulate cortex: An update. *Trends in Cognitive Science*, 8, 539–546.

Bromberg-Marting, E. S., Matsumoto, M., & Hikosaka, O. (2010). Dopamine in motivational control: Rewarding, aversive and alerting. *Neuron*, 68, 815–832.

Collins, A. G., & Frank, M. J. (2013). Cognitive control over learning: Creating, clustering and generalizing task-set structure. *Psychological Review*, 120, 190–229.

Corbit, L. H., Muir, J. L., & Balleine, B. W. (2001). The role of the nucleus accumbens in instrumental conditioning: Evidence of a functional dissociation between accumbens core and shell. *Journal of Neuroscience*, 21, 3251–3260.

Cossell, L., Iacaruso, M. F., Muir, D. R., Houlton, R., Sader, E. N., Ko, H., … Mrsc-Flogel, T. D. (2015). Functional organization of excitatory synaptic strength in primary visual cortex. *Nature*, 518, 399–403.

Couteure, E., & Killcross, S. (2003). Inactivation of the infralimbic prefrontal cortex reinstates goal-directed responding in overtrained rats. *Behavioural Brain Research*, 146, 167–174.

Cushman, F., & Morris, A. (2015). Habitual control of goal selection in humans. *Proceedings of the National Academy of Sciences of the United States of America*, 112, 13817–13822.

Daw, N. D., Niv, Y., & Dayan, P. (2005). Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioral control. *Nature Neuroscience*, 8(12), 1704–1711.

Dayan, P., & Abbot, L. (2001). *Theoretical neuroscience: computational and mathematical modeling of neural systems*. Cambridge, MA: MIT Press.

Dezfouli, A., & Balleine, B. W. (2012). Habits, action sequences and reinforcement learning. *European Journal of Neuroscience*, 35, 1036–1051.

Dezfouli, A., & Balleine, B. W. (2013). Actions, action sequences and habits: Evidence that goal-directed and habitual action control are hierarchically organized. *Plos Computational Biology*, 9, 1–14.

Dhamala, M., Rangarajan, G., & Ding, M. (2008). Analyzing information flow in brain networks with nonparametric granger causality. *NeuroImage*, 41, 354–362.

Doll, B. B., Simon, D. A., & Daw, N. D. (2012). The ubiquity of model-based reinforcement learning. *Current Opinion in Neurobiology*, 12, 1075–1081.

Durstewitz, D., Vitoz, N. M., Floresco, S. B., & Seamans, J. K. (2010). Abrupt transitions between prefrontal neural ensemble states accompany behavioral transitions during rule learning. *Neuron*, 66, 438–448.

Engelhard, B., Finkelstein, J., Cox, J., Fleming, W., Jang, H. J., Ornelas, S., … Witten, I. B. (2019). Specialized coding of sensory, motor and cognitive variables in VTA dopamine neurons. *Nature*, 570, 509–513.

Frank, M. J. (2011). Computational models of motivated action selection in corticostratal circuits. *Current Opinion in Neurobiology*, 21, 381–386.

Frank, M. J., & Badre, D. (2012). Mechanisms of hierarchical reinforcement learning in corticostratal circuits 1: Computational analysis. *Cerebral Cortex*, 22, 509–523.

Fregnan, Y., Panacceau, M., Rene, A., Huguet, N., Marre, O., Levy, M., & Shulz, D. E. (2010). A re-examination of Hebbian-covariance rules and spike timing-dependent plasticity in cat visual cortex in vivo. *Frontiers in Synthetic Neuroscience*, 2, 1–21.

Garthe, A., Roeder, I., & Kempermann, G. (2016). Mice in an enriched environment learn more flexibly because of adult hippocampal neurogenesis. *Hippocampus*, 26, 261–271.

Gershman, S. J., Moore, C. D., Todd, M. T., & Normal, K. A. (2012). The successor representation and temporal context. *Neural Computation*, 24, 1553–1568.

Gonner, L., Vitay, J., & Hamker, F. H. (2017). Predictive place-cell sequences for goal-finding emerge from goal memory and the cognitive map: A computational model. *Frontiers in Computational Neuroscience*, 11, 1–19.

Groenewegen, H., Wouterlood, F., & Uylings, H. (2017). Organization of prefrontal-striatal connections. In H. Steiner, & K. Tseng (Eds.), *Handbook of Basal Ganglia structure and function*, chapter 21 (pp. 433–450). London: Academic Press.

Groenewegen, H. J., Wright, C. I., Beijer, A. V., & Voorn, P. (1999). Convergence and segregation of ventral striatal inputs and outputs. *Annals of the New York Academy of Sciences*, 877, 49–63.

Groenewegen, H. J., Wright, C. I., & Uylings, H. B. (1997). The anatomical relationships of the prefrontal cortex with limbic structures and the basal ganglia. *Journal of Psychopharmacology*, 11, 99–106.

Haber, S. N. (2003). The primate basal ganglia: Parallel and integrative networks. *Journal of Chemical Neuroanatomy*, 26, 317–330.

Haber, S. N. (2016). Corticostratal circuity. *Dialogues in Clinical Neuroscience*, 18, 7–21.

Haber, S. N., & Knutson, B. (2010). The reward circuit: Linking primate anatomy and human imaging. *Neuropsychopharmacology*, 35, 4–26.

Helie, S., Chakravarthy, S., & Moustafa, A. A. (2013). Exploring the cognitive and motor functions of the basal ganglia: An integrative review of computational cognitive neuroscience models. *Frontiers in Computational Neuroscience*, 7, 1–16.

Isoda, M., & Hikosaka, O. (2007). Switching from automatic to controlled action by monkey medial frontal cortex. *Nature Neuroscience*, 10, 240–248.

Ito, M., & Doya, K. (2011). Multiple representations and algorithms for reinforcement learning in the cortico-basal ganglia circuit. *Current Opinion in Neurobiology*, 21, 368–373.

Joel, D., & Weiner, I. (1994). The organization of the basal ganglia-thalamocortical circuits: Open interconnected rather than closed segregated. *Neuroscience*, 63, 363–379.
Smith, K. S., Virkud, A., Deisseroth, K., & Graybiel, A. M. (2012). Reversible online control of habitual behavior by optogenetic perturbation of medial prefrontal cortex. *Cerebral Cortex, 109*, 18932–18937.

Stefani, M. R., & Moghaddam, B. (2006). Rule learning and reward contingency are associated with dissociable patterns of dopamine activation in the rat prefrontal cortex, nucleus accumbens, and dorsal striatum. *Journal of Neuroscience, 26*, 8810–8818.

Sutton, R. S., & Barto, A. G. (2018). *Reinforcement Learning: An Introduction*. Cambridge, MA: MIT Press.

Tehovnik, E. J., & Sommer, M. A. (1997). Effective spread and time-course of neural inactivation caused by lidocaine injection in monkey cerebral cortex. *Journal of Neuroscience Methods, 74*, 17–26.

Thurley, K., & Ayaz, A. (2017). Virtual reality systems for rodents. *Current Zoology, 63*, 109–119.

Topalidou, M., Kase, D., Boraud, T., & Rougier, N. P. (2018). A computational model of dual competition between the basal ganglia and the cortex. *eNeuro, 5*, 1–17.

Turner, C. A., Lewis, M. H., & King, M. A. (2003). Environmental enrichment: Effects on stereotyped behavior and dendritic morphology. *Developmental Psychobiology: The Journal of the International Society for Developmental Psychobiology, 10*, 20–27.

Turner, C. A., Yang, M. C., & Lewis, M. H. (2002). Environmental enrichment: Effects on stereotyped behavior and regional neuronal metabolic activity. *Brain Research, 938*, 15–21.

Villagrasa, F., Baladron, J., Vitay, J., Schroll, H., Antzoulatos, E. G., Miller, E. K., & Hamker, F. H. (2018). On the role of the cortex-basal ganglia interactions for category learning: A neuro computational approach. *Journal of Neuroscience, 38*, 9551–9562.

Vitay, J., Dinkelbach, H. U., & Hamker, F. H. (2015). ANNarchy: A code generation approach to neural simulations on parallel hardware. *Frontiers in Neuroinformatics, 9*, 1–10.

Wilton, J., & Hamker, F. H. (2009). Efficient coding correlates with spatial frequency tuning in a model of V1 receptive field organization. *Visual Neuroscience, 26*, 21–34.

Yin, H. H. (2014). Cortico-basal ganglia networks and the neural substrates of actions. In L. M. Chalupa, & J. S. Werner (Eds.), *Neurobiology of alcohol dependence*, chapter, vol. 18 (pp. 260–278). London: Academic Press.

Yin, H. H. (2016). The basal ganglia and hierarchical control in voluntary behavior. In J.-J. Soghomonian (Ed.), *The Basal Ganglia novel perspectives on motor and cognitive functions*, chapter 20, (pp. 513–566). Basel, Switzerland: Springer.

Yin, H. H., & Knowlton, B. J. (2006). The role of the basal ganglia in habit formation. *Nature Reviews Neuroscience, 7*, 464–476.

Yin, H. H., Knowlton, B. J., & Balleine, B. W. (2004). Lesions of dorsolateral striatum preserve outcome expectancy but disrupt habit formation in instrumental learning. *European Journal of Neuroscience, 19*, 181–189.

Yin, H. H., Mulcare, S. P., Hilario, M. R. F., Clouse, E., Holloway, T., Davis, M. L., ... Costa, R. M. (2009). Dynamic reorganization of striatal circuits during acquisition and consolidation of a skill. *Exo Brain Research, 146*, 333–341.

Yin, H. H., Ostlund, S. B., Knowlton, B. J., & Balleine, B. W. (2005). The role of the dorsomedial striatum in instrumental conditioning. *European Journal of Neuroscience, 22*, 513–523.

How to cite this article: Baladron J, Hamker FH. Habit learning in hierarchical cortex–basal ganglia loops. *Eur J Neurosci*. 2020;52:4613–4638. https://doi.org/10.1111/ejn.14730