Intrinsically Motivated Action-Outcome Learning and Goal-Based Action Recall: A System-Level Bio-Constrained Computational Model

Gianluca Baldassarre[1], Francesco Mannella[1], Vincenzo G. Fiore[1], Peter Redgrave[2], Kevin Gurney[2], Marco Mirolli[1]

[1]Laboratory of Computational Embodied Neuroscience,
Istituto di Scienze e Tecnologie della Cognizione,
Consiglio Nazionale delle Ricerche (LOCEN-ISTC-CNR),
Via San Martino della Battaglia 44, I-00185 Roma, Italy
{gianluca.baldassare,francesco.mannella,vincenzo.fiore,
valerio.sperati,daniele.caligiore,marco.mirolli}@istc.cnr.it
[2]Adaptive Behaviour Research Group,
Department of Psychology, University of Sheffield, S10 2TP, UK
{p.redgrave,k.gurney}@sheffield.ac.uk

Abstract

Reinforcement (trial-and-error) learning in animals is driven by a multitude of processes. Most animals have evolved several sophisticated systems of 'extrinsic motivations' (EMs) that guide them to acquire behaviours allowing them to maintain their bodies, defend against threat, and reproduce. Animals have also evolved various systems of 'intrinsic motivations' (IMs) that allow them to acquire actions in the absence of extrinsic rewards. These actions are used later to pursue such rewards when they become available. Intrinsic motivation has been studied in Psychology for many decades and its biological substrate is now being elucidated by neuroscientists. In the last two decades, investigators in computational modelling, robotics and machine learning have proposed various mechanisms that capture certain aspects of IMs. However, we still lack models of IMs that attempt to integrate all key aspects of intrinsically motivated learning and behaviour while taking into account the relevant neurobiological constraints. This paper proposes a bio-constrained system-level model that contributes a major step towards this integration. The model focusses on three processes related to IMs and on the neural mechanisms underlying them: (a) the acquisition of action-outcome associations (internal models of the agent-environment interaction) driven by phasic dopamine signals caused by sudden, unexpected changes in the environment; (b) the transient focussing of visual gaze and actions on salient portions of the environment; (c) the subsequent recall of actions to pursue extrinsic rewards based on goal-directed reactivation of the representations of their outcomes. The tests of the model, including a series of selective lesions, show how the focussing processes lead to a faster learning of action-outcome associations, and how these associations can be recruited for accomplishing goal-directed behaviours. The model, together with the background knowl-

edge reviewed in the paper, represents a framework that can be used to guide the design and interpretation of empirical experiments on IMs, and to computationally validate and further develop theories on them.

Keywords: Intrinsic motivations; trial-and-error learning; attention; superior colliculus; dopamine; repetition bias; striato-cortical loops; basal ganglia selection; parietal, premotor, prefrontal cortex.

1 Introduction

Most organisms are endowed with complex systems of *extrinsic motivations* (EMs) that drive the execution and the acquisition of behaviours that serve homeostatic regulation. This enhances their biological fitness by allowing them, for example, to escape predators, seek food and water, and reproduce. One of the hallmarks of mammals, and in particular primates, is their capacity to learn on the basis of *intrinsic motivations* (IMs). The notion of IM was initially developed because the classical theories of instrumental learning and drives (e.g., Skinner, 1938; Hull, 1943) fell short in their ability to explain some empirical findings; for example why monkeys spontaneously engage in puzzles (Harlow, 1950), or why rats can be instrumentally trained with an apparently-neutral stimulus (such as the sudden onset of a light) to perform an action without an extrinsic reward (e.g., food) (Kish, 1955). Berlyne (1966) systematically studied the properties of certain stimuli, traditionally not considered to be reinforcing, that can trigger spontaneous exploration: such stimuli tend to be complex, unexpected, or in general 'surprising'. Later, other researchers, giving a stronger emphasis on the relation between IMs and *action*, proposed that an important aspect of these motivations is the capacity of animals to impact the world with their own actions (e.g., based on the concept of "effectance", White, 1959).

Some recent computationally grounded work has proposed a theoretical systematisation of IMs. In particular, Oudeyer and Kaplan (2007) have clarified the existence of two classes of computational mechanisms to implement IMs; those based on measures of knowledge on stimuli (predictability and novelty) lead to *knowledge-based IMs (KB-IMs)*, whereas those based on measures of action acquisition lead to *competence-based IMs (CB-IMs)*. Mirolli and Baldassarre (inpr) have clarified how all these mechanisms serve the ultimate function of action acquisition and performance, but also that they can do this by implementing two distinct sub-functions, namely, the acquisition of knowledge or the acquisition of competence. In this respect, *both* KB-IMs and CB-IMs can be used for either sub-function (e.g., as further discussed in Sec. 4, the model presented here exploits a KB-IM mechanism to drive the acquisition of competence). Singh et al. (2010) and Schembri et al. (2007c) have proposed evolutionary computational models to explain the adaptive origin of IMs and their close relation to EM. Related to this, Baldassarre (2011) has proposed that IMs have the adaptive function of driving the acquisition of actions when rewards produced by EM (e.g., related to food and sex) are temporally distal, or would require the acquisition of overly complex behaviours. At a later stage, the actions acquired through IMs can then be readily recalled or assembled to achieve extrinsic rewards when these become available. Further, related to the model proposed here, Baldassarre (2011) has started to distinguish EM and IMs on the basis of the differences between the brain mechanisms underlying them.

Recently, neuroscience has started to propose theories related to IMs. For example, it has been shown how hippocampus responds to novel stimuli (or novel spatio-temporal relations between familiar stimuli) thereby generating a dopaminergic learning signal that might drive the formation of new memories (Lisman and Grace, 2005; Kumaran and Maguire, 2007). Research on locus

coeruleus has shown how this nucleus produces noradrenaline when environmental predictions are violated, and how this might drive learning processes within its target structures (Sara et al., 1994; Sara, 2009; cf. Yu and Dayan, 2003, for a related model). Dopamine (DA) has also been shown to be related to IMs, for example novel stimuli cause its release (Schultz, 1998; Kakade and Dayan, 2002 show how a popular reinforcement-learning algorithm – TD(0), see Sutton and Barto, 1998 – captures this data if enhanced with 'exploration bonuses', transient rewards or higher evaluations for novel stimuli).

One of the most comprehensive neuroscientific theories relating DA and IMs has been proposed by Redgrave and Gurney (2006) and Redgrave et al. (2011). This theory highlights some important mechanisms underlying IMs that are pivotal for the model proposed here. Key structures in this theory are: the basal ganglia (BG) – a group of subcortical nuclei important for action selection and reinforcement learning in operant conditioning; the striatum (Str) – the major input to BG; superior colliculus (SC) – a subcortical nucleus receiving input from the retina and important for controlling eye movements. According to Redgrave and Gurney (2006) sudden unexpected events activate the superior colliculus that, in turn, triggers phasic responses in midbrain dopamine neurons, with bursts whose amplitude diminishes as the stimulus becomes familiar. These dopamine neurons innervate striatum, and the phasic release of dopamine here causes plasticity which facilitates cortico-striatal transmission. This results in the most recently selected actions being more likely to be selected again, so that there is a tendency to repeat the actions that caused the phasic event. This phenomenon is referred to as *repetition-bias* and we conceive of it as an example of IM since the external stimulus triggering learning (e.g. a sudden light onset) is not extrinsically rewarding. Bolado-Gomez et al. (2009) have shown how a biologically plausible learning rule could produce repetition bias in a behaving agent, and Gurney et al. (2009) have demostrated that the learning rules required for action acquition are consistent with recent *in vitro* data. The repetition of the action and its consequences causes representations of the action and its outcome to be repeatedly presented at brain structures (including cortical areas) which can so form associations between them. In this way, *internal models* of the *action-outcome* contingencies can be stored. Once acquired on the basis of IMs, these models allow actions potentially leading to specific outcomes to be recalled through the activation of the neural representations of such outcomes, thereby allowing *goal-directed behaviour* (hence we will refer to internally re-activated neural representations of desired outcomes as *goals*). This theory has recently been articulated in more detail by Gurney et al. (2012) where the notions of action selection, prediction errors, internal models, etc., have been given formal ontological definitions.

The contributions reviewed above represent important advancements for our understanding of IMs and their relations to EM. However, we still lack a complete, fully specified, systems-level model which integrates all aspects of the theory of Redgrave and Gurney (2006), including the learning of internal model associations and their subsequent recall in goal-directed behaviour. This work proposes a model that fills this gap in our knowledge. In particular, the model investigates the following aspects of behaviour and their possible underlying brain mechanisms. First, the role of repetition bias, induced by phasic and unexpected environmental changes, and its ability to facilitate the acquisition of action-outcome associations in cortico-cortical neural pathways involving prefrontal, motor, and parietal cortex. Second, the later recall of actions directed to pursue biologically valuable effects based on the activation of the representations of desired outcomes (goals). Notice that although most of these aspects have been empirically investigated and theoretically discussed in previous works, the model presented here is the first to: (a) specifying all of them to a detail that allows their computational implementation; (b) integrating them into a complete, autonomously

functioning system; (c) doing so while obeying biological constraints at the macro-architectural (system) level.

In order to accommodate these mechanisms, the architectural and functional scope of the model is necessarily broad. The model was therefore developed and constrained with biological data at the *system-level*, that is at the level of the functions played by the various brain nuclei and subsystems, while deferring to future work the goal of introducing stronger constraints at the micro-circuit and physiological level. Further, some computations which are more tangential to the overall hypothesis (e.g., the details of oculomotor control) are subsumed into simple functional elements rather than implemented in a biologically plausible ways.

The rest of the paper is organised as follows. Sec. 2.1 illustrates the task used to test the model: this task is similar to a behavioural experiment on IMs with monkeys and children (Taffoni et al., inpr). Sec. 2.2 gives an overview of the model components and then describes how they work at a functional level during the learning and test phases. Sec. 2.3 describes in more detail the biological constraints used to design the model. Sec. 2.4 presents the computational mechanisms used to implement the model in detail. Sec. 3 presents the results of tests of the model, both in its complete, functioning form, and after several lesions designed to dissect the correspondence between mechanism and functions: the results of these experiments represent predictions of the model that might be tested in future experiments. Finally, Sec. 4 discusses the results and the key features of the model both from a biological and a computational perspective.

The paper's remit is extensive as it aims to: (i) integrate biological and computational aspects of IM; (ii) present a specific model of IM; (iii) a general framework to carry out other investigations on IM. For this reason, to aid navigation through the text, several sections are readable independently of others. The reader with focused interests can therefore read a sub-set of the paper without loss of narrative. In particular, the reader interested in the biological aspects related to IMs can read the model overview (Sec. 2.2), the biological framework (Sec. 2.3), and then jump to the discussion of the biological issues (Sec. 4.1). The reader interested in the computational details of the model might instead browse the overview of the model (Sec. 2.2), read thoroughly its computational details (Sec. 2.4) and the results (sections and 3), and only then access the discussion of the computational aspects (Sec. 4.2). Finally, the reader interested only in the computational results of the model might only read the model overview (Sec. 2.2) and the results (Sec. 3).

2 Methods

2.1 The target behavioural task

The task is an abstraction of an experiment described in Taffoni et al. (2012) and (Taffoni et al., inpr) specifically designed to test theories on IMs, and makes an ideal test bed for the model. The apparatus has been adapted for use by monkeys and children, as shown in Figure 1.

The apparatus is formed by a working plane having three buttons that can be pressed, and a vertical plane in which there are small embedded boxes that can be opened with the buttons and can so deliver a reward. In this plane are also set three sets of lights and a loudspeaker for additional feedback.

The experiment is divided into two phases: a *learning* phase and a *test* phase. There are control and experimental groups of participants. In the learning phase, if a participant of the control group presses a certain button, a certain set of lights and sounds are turned on. If a participant of the experimental group presses a button, the board produces the same effects with the additional

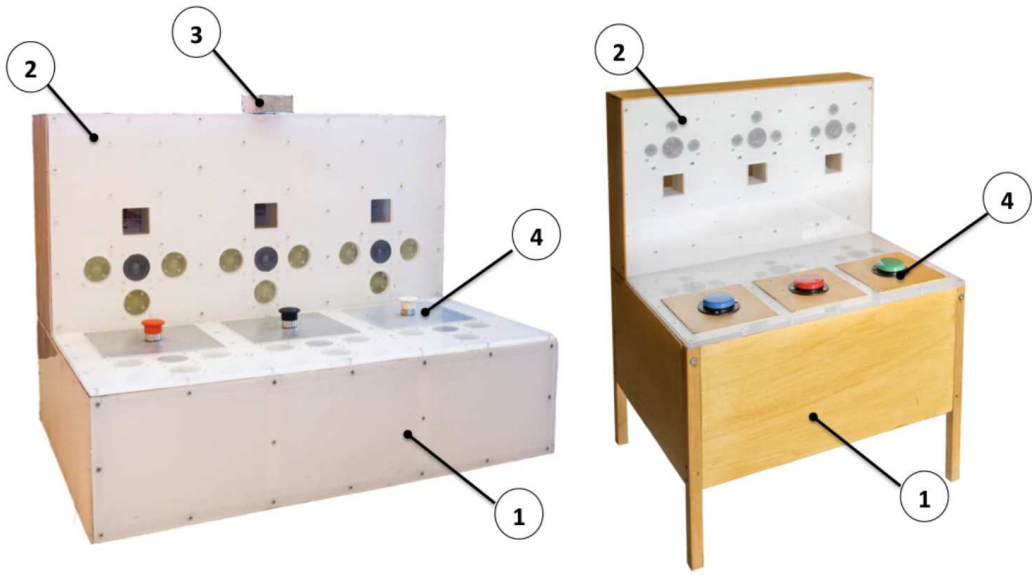


Figure 1: The experimental apparatus that was used with monkeys (left) and children (right) to run the target experiments. The apparatus is formed by a working plane (1) and a feedback plane (2) (3: embedded camera used for monitoring the monkeys). The working plane (4) contains mechatronic objects that can be manipulated (here simple buttons that can be pressed). The feedback plane contains lights, loudspeakers, and boxes that can open automatically: these can produce outcomes caused by actions executed on the mechatronic objects.

opening of a corresponding box. In this phase, no extrinsic reward is given and the participant should be driven to explore the board by IMs.

In the test phase, which is identical for the control and experimental group, a reward is set inside one of the three boxes. For monkeys, this was a peanut, for children a sticker (a reward token used extensively by developmental psychologists). The reward is visible to the participants as the boxes have a transparent cover, so it should motivate them to recall the action required to open the box acquired during the learning phase. During the test phase, the time taken for the retrieval of the reward is measured. The results of pilot experiments with children (Taffoni et al., 2012, Taffoni et al., inpr) indicate that subjects in the experimental group tend to retrieve the rewards faster than the control group. These preliminary results can be explained by the experimental hypothesis under which the participants' exploration during the learning phase allows them to acquire the action-outcome associations (press button x $\leftrightarrow$ box y opens) required in the subsequent test phase to retrieve the reward.

This experiment is relevant for the investigation of IMs for the following reasons. The experiment does not manipulate IMs. Indeed, IMs are present in both the control and experimental groups in the same way. The idea of the experiment is instead to demonstrate that IMs allow the agents to learn things that can be later exploited to gain extrinsically rewarded outcomes. To this purpose, the experiment design assures that: (a) In the first training phase of the experiment there are

clearly no EM involved, so any knowledge or competence acquisition is based on IMs; (b) In this phase, the experiment manipulated what could be learned by the two groups (opening of the box or not); (c) In the second test phase the knowledge/competence eventually acquired in the first phase can be exploited to improve performance. The hypothesis of the experiment was that the second phase would have shown a different performance of the control and experimental groups, so supporting the idea that IMs can indeed have the function of acquiring knowledge/competence that can be later exploited for better achieving extrinsic rewards.

We now describe the simulated schedule used to test the model. During the learning phase, which lasted 60 mins of simulated time, the model freely explored the board and learned action-outcome associations based on IMs. During the test phase, lasting 6 mins, the model was tested to see if it was capable of recalling the required actions on the basis of the activation of its internal representations of the action outcomes. In particular, each outcome representation (box-open) was activated for 2 mins and the model had to make repeated sequences of actions with this goal in place. If a box opened it stayed open for 2 sec before 'closing'. The model did not undergo any reset or re-initialisation.

The inputs and actions of the model were encoded at a rather coarse level of granularity given the focus of this research on high-level aspects of cognition. In particular in the model the actions represented whole movement sequences such as 'look at button x', 'look at box x', 'press x', etc. This choice is in line with the empirical experiments with monkeys and children which clearly indicate that, when these participants face the board experiment, they already posses a rich repertoire of orienting and manipulative actions acquired in previous life experiences. In particular, they probe the board by executing quite complex actions (such as the ones mentioned above) and seem to perform a kind of 'motor-babbling' (von Hofsten, 1982) albeit at the *complete action* level rather than at the fine movement level. The inputs to the model were chosen at a similar coarse granularity and so they represented entire objects such as the buttons and boxes, or the context corresponding to the whole experiment.

2.2 Overview of the model

Figure 2 shows the components of the model architecture, highlighting their main function[1] and their possible biological correspondents. A more detailed illustration of the architecture is given in Figure 3, and a fuller account of its biological basis in Sec. 2.3; a list of acronyms referring to brain areas is given in table 2 in the Appendix. The core of the model is formed by three coupled components, corresponding to three BG-cortical loops: one loop selects the arm actions ('arm loop'), one selects the eye gaze ('oculomotor loop'), and one selects the goals to pursue during the test phase ('goal loop').

The oculomotor loop has a constant input representing the overall context of being situated in the experiment, and can select a saccade to a point from among six possible spatial locations – the three buttons and the three boxes. The arm loop receives six different possible input patterns corresponding to the possible perceived objects – again the three buttons and three boxes. On this basis the arm loop can select among three possible actions: a 'reach to and press the looked-at object' action, that when performed on a button opens the corresponding box, and two inconsequential ('dummy') actions introduced to test the learning capabilities of the arm loop (e.g. 'reach and

[1]Note that we describe each component of the model as implementing a 'main function' for ease of explanation: in fact, each component implements specific computational processes and its function emerges from these processes and the interaction of the component with other components of the model with which it interacts.

point at, but don't press' and 'reach and wave' – the precise definitions are not important). During the learning phase, when an action causes a box to open, a representation of this outcome occurs within the goal loop, allowing the formation of associations between such representations and the representations of the actions within the arm and oculomotor loops. During the test phase, the goal loop receives an external ('hardwired') input that abstracts the values assigned to the different possible outcomes by sub-cortical brain systems: this causes the selection of one particular goal.

Another key component of the model is the SC that is activated by sudden unexpected luminance changes and which can, as a consequence, generate a phasic response in midbrain dopaminergic areas (Comoli et al., 2003). This signal drives a learning process involving striatal afferent connections of the arm and oculomotor loop. It is known that the amplitude of the phasic dopamine signal declines as the stimulus becomes predicted (see for example, Schultz, 2010). Recently Shah and Gurney (2011) have shown that, computationally, this phenomenon is required to prevent unlearning and can, under some circumstances, enable an optimal outcome without a specific cost function. Here, we do not model the prediction mechanism as such, but simply model the decay, or inhibition, of the dopamine response phenomenologically. Thus, we force the amplitude of phasic dopamine to decline when a salient event happening at a specific place, for example the opening of a box, is experienced several times; we refer to this mechanism as the 'dopamine inhibitor'.

Finally, we model the initiation of reflexive saccades in response to phasic, peripheral stimuli. Such a response is required in order to cause saccades to phasic events such as the opening of one of the boxes. Anatomically, this process involves a subcortical loop formed by BG and SC (a full model of integrated subcortical and cortical gaze control has been recently given by Cope and Gurney, 2011). However, the precise mechanisms at work here are not part of our primary interest and so we model reflexive saccades phenomenologically by simply overriding the output of the oculomotor loop when required.

During the learning phase, the model operates as follows. The oculomotor and arm loops initially select actions in a random fashion. In particular, the eye foveates one of the six visually salient positions of the board (three button, three boxes) and the arm performs one of the three available actions (the 'reach and press' action, and the two inconsequential dummy actions). Note that there are 18 possible combinations of the oculomotor and arm loop actions. Occasionally, a combination of these will occur which causes a box to open (looking at a button and reaching/pressing it). The initially unexpected environmental event of box-opening activates the SC that causes a phasic DA burst. When such an event is experienced several times, the DA inhibitor progressively attenuates the corresponding DA signal. Further, immediately after the event, the reflexive saccade system operates to drive the system to foveate to the portion of space where the change took place (the opening of a box).

The perception of the environmental change, with the consequent DA signal, triggers two learning processes, one involving the cortex and one involving the striatum. The learning process involving cortex modifies the connections projecting from the goal loop to the arm loop and to the oculomotor loop. This process is based on a Hebbian learning rule involving DA (see Sec. 2.4) and forms associations between the outcome currently activated in the goal loop (e.g., 'box x opens'), and the actions just performed, i.e. a combination of the 'saccade to button x' eye action and the 'reach-and-press the looked-at object' arm action encoded respectively in the oculomotor and arm loops.

In the striatal learning process, phasic DA reaches the striatal afferent connections of the arm and oculomotor loops and drives plasticity based on a 3-factor learning rule (Reynolds and Wickens, 2002) (see Sec. 2.4). This results in a strengthening of the connections between the seen object and

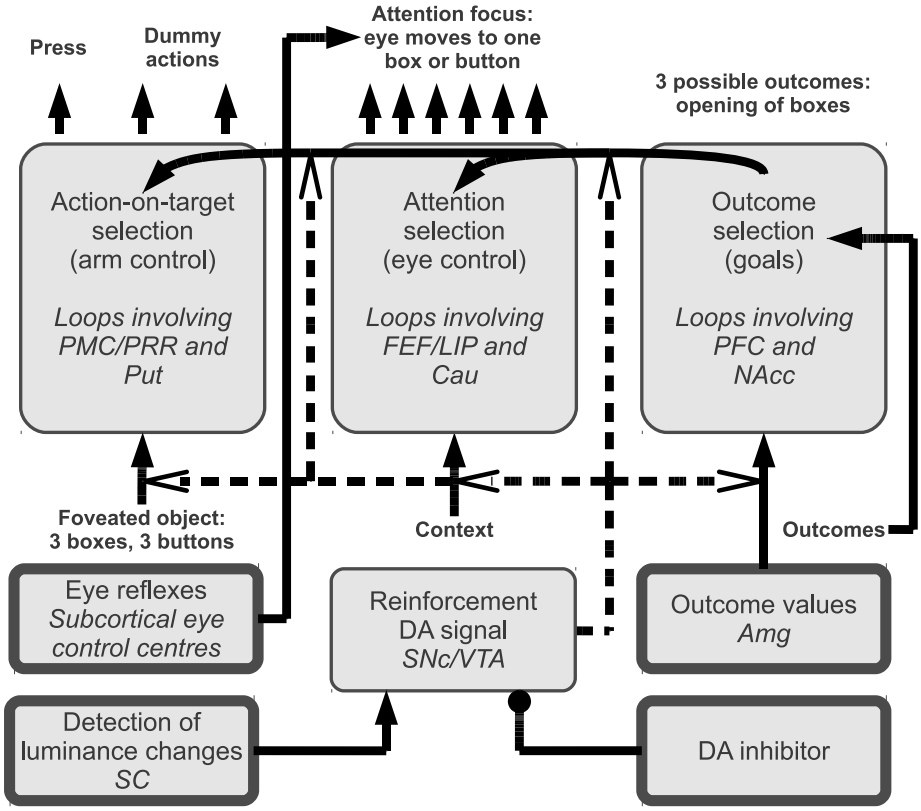


Figure 2: Functional representation of the model showing its main components and their macro-circuit connectivity. Arrow heads represent excitatory connections, whereas circle-heads represent inhibitory connections. Dashed lines represent the dopamine learning signals. Boxes with a bold border are phenomenologically modelled ('hardwired') components. See table 2 for the acronyms used in the figure.

the performed action within the arm loop, and between the 'context' and the eye saccade within the oculomotor loop. In this way the system can learn to look at a specific portion of space and to select the correct arm action ('reach and press') when it sees a button so as to cause an interesting change in the environment ('opening of a box'). When the DA signal decreases due to the action of the 'DA inhibitor' (subsuming, in our model, a prediction process), the striatal afferent connection weights undergo a spontaneous decay. This leads the oculomotor loop to recommence exploration of the environment, and the arm loop to select other actions so that the system can discover new unpredicted events that its actions can cause.

The test phase aims to show how the action-outcome associations acquired in the intrinsically motivated learning phase may be subsequently recruited to allow the system to pursue rewarding goals (e.g., food). During the test phase we manually activate each outcome representation encoding a certain box-opening within the goal loop for a certain time (see Sec. 2.4) and record the successful

acquisition of the external rewards. If the system is capable of learning suitable representations of the possible action-outcome contingencies based on IMs, there should be a higher rate of box-openings in the experimental group than in the control group. The mechanism here is that, if suitable cortico-cortical connections are formed in the learning phase, then when a goal is activated the activity propagates from the goal loop to the oculomotor and arm loops causing respectively a saccade towards the button that can open the corresponding box and a 'reach and press' action at that button.

2.3 The biology underpinning the model

This section presents the biological constraints that has been used to select the model components and their function illustrated in figure 2 and, in some cases, the internal micro-architecture and functioning illustrated in Sec. 2.4. In so doing, we will review a the biological evidence that we think is relevant to investigate the IM phenomena targeted here. Table 1 summarises the brain sub-systems relevant for the phenomena investigated here, the main functions they might implement, and some references used as sources of this information. Note that not all these constraints find their way directly into the current model, and in some cases some biological functions are incorporated in the model at an abstract, phenomenological level. Working with these multiple levels of description is a principled methodology if done in such a way as to preserve representation semantics between the levels (Gurney, 2009; Gurney et al., 2004). Notwithstanding this selective use of constraints, we present this more comprehensive review of the biology as we believe it lays the foundation for the future development of the model (see also Sec.4).

Table 1: Key references on the main functions ascribed to the brain components and neural systems forming the model (components in brackets are not explicitly modelled here). A table of the acronyms can be found in the Appendix.

Brain area	**Function**	**References**
Cortical bottom-up neural streams		
(VC→PC→PMC/PFC)	Dorsal neural pathway	Jeannerod (1999) Luppino and Rizzolatti (2000) Simon et al. (2002) Rizzolatti and Matelli (2003) Cisek and Kalaska (2010)
(VC→PRR→PMCd)	Affordances, arm control	Wise et al. (1997)
(VC→LIP→FEF)	Voluntary eye control	Snyder et al. (2000)
VC→ITC→PFC	Object recognition, context	Grill-Spector and Malach (2004)
Sub-cortical/cortical top-down neural pathways		
(Amg→PFC)	Assigns 'extrinsic' values to objects and events	Wallis (2007)
PFC→PMC/PRR and · · · PFC→FEF/LIP	Biases affordance and action selection	Miller and Cohen (2001) Fuster (2008) Yeterian et al. (2011)
VTA→Cortex	Dopamine based learning processes	Otani et al. (2003)

Continued on next page

Brain area	**Function**	**References**
		Huang et al. (2004)
BG-cortical loops		
BG↔cortex	Macro loops,	Alexander et al. (1986)
	channel organisation,	Chevalier and Deniau (1990)
	selection of cognitive contents	Redgrave et al. (1999)
		Haber (2003)
		Romanelli et al. (2005)
		Yin and Knowlton (2006)
		Ashby et al. (2010)
Input→Put, Input→Cau	Trial-and-error learning,	Redgrave et al. (2011)
	action repetition bias	Berridge and Robinson (1998)
		Berridge et al. (2005)
Amg→NAcc	Goal-selection based on values	Cardinal et al. (2002)
		Pennartz et al. (inpr)
Put→GPi→Th→PMC/PRR	Selection of arm movements	Jaeger et al. (1993)
Cau→SNr→Th→FEF/LIP	Selection of eye movements	Hikosaka et al. (2000)
NAcc→SNr→Th→PFC	Selection of goals	Middleton and Strick (2002)
		Haber (2003)
Cortex→Str	Trial-and-error learning,	Houk et al. (1995)
	LTP and LTD processes	Wickens (2009)
Others		
SC	Generates DA learning signals	Redgrave and Gurney (2006)
	with sudden unexpected events	Comoli et al. (2003)
Subcortical eye centers	Reflex eye movements	Hikosaka et al. (2000)
Inhibitor	Progressively inhibits DA	Smith et al. (2004)
		Balcita-Pedicino et al. (2011)
(Amg)	Assigns value to goals	Pitkänen et al. (1997)
		Cardinal et al. (2002)
		Balleine et al. (2003)
		Mirolli et al. (2010)
(Hip)	Responds to novel stimuli and	Lisman and Grace (2005)
	novel spatial/temporal relations	Kumaran and Maguire (2007)
(LC)	Signals violations of expectations	Sara et al. (1994); Sara (2009)

As mentioned in Sec. 2.2, the core of the model is formed by three main components representing three *BG-cortical loops*: these loops have the role of selecting arm actions, eye gaze, and goals. They also implement important reinforcement learning processes contingent on phasic dopamine (Reynolds and Wickens, 2002). The basal gangia (BG) are a set of sub-cortical nuclei tightly linked to associative and frontal cortex via re-entrant connections (Alexander et al., 1986; Houk et al., 1995; Hikosaka, 1998; Redgrave et al., 2011). The striatum is the major input gateway of the BG and cortex sends important efferent connections to the striatum mainly from layer V. Cortex also receives afferent connections from BG via the thalamus (Th) mainly within layer IV (Kandel et al., 2000; Shepherd and Grillner, 2010). Within BG the signals are processed via a double inhibition mechanism involving the striatum as a first stage, and the internal globus pallidus (GPi) and

substantia nigra pars reticulata (SNr) as a second stage (both stages involve GABAergic efferent connections; see Chevalier and Deniau, 1990).

The striatum and GPi/SNr implement the so-called *direct pathway* of BG. The inhibitory action of GPi can also be augmented by the diffuse excitatory efferent connections it receives from the subthalamic nucleus (STN), itself receiving input from the cortex. The cortico-subthalamo-pallidal pathway is sometimes referred to as the *hyperdirect pathway* (Nambu et al., 2002). The BG also involve a third indirect circuit, the *indirect pathway*, incorporating the external globus pallidus not considered in the model presented here. The pathways of BG-cortical loops tend to form partially segregated *channels* as their different portions (striatum, GPi/SNr, and STN), and the thalamic regions to which they project, generally preserve the topological organisation of the cortex to which they are connected (Haber, 2003; Romanelli et al., 2005). The organisation into channels of the direct pathway (suitably modulated by the indirect and hyperdirect pathways), has led to wide agreement that the BG are well suited to select secondary perceptual representations, actions, and other more complex cognitive contents when animals face selection problems in these domains (Houk et al., 1995; Redgrave et al., 1999; Joel et al., 2002). The selection processes within the BG might be further strengthened via inhibitory connections internal to the Th (Crabtree and Isaac, 2002).

The BG are also a key brain structure important for reinforcement (trial-and-error) learning processes (Houk et al., 1995; Doya, 1999; Brown et al., 1999; Joel et al., 2002). In particular, the cortical synaptic contacts to striatal neurons have been shown to undergo LTP and LTD modulated by dopamine (Reynolds and Wickens, 2002; Wickens et al., 2003; Calabresi et al., 2007; Wickens, 2009). An aspect of these learning processes important for the model is that the associative striatum is particularly important for task acquisition and/or performance during early stages of learning, while sensorimotor striatum responds more strongly after it has become habitual or automatised (see Ashby et al., 2010, for a review). This process is finessed in regards to whether the habitual action is simple or is composed of a sequence of simpler actions. In the former case, extinction of sensorimotor striatal activity may occur even in habitual tasks (e.g. Carelli et al., 1997).

The organisation of BG-cortical loops around partially segregated functional channels also tends to repeat at a higher level of organisation in terms of the *classes* of action being selected. Thus, different BG-cortical loops tend to form whole systems that play partially distinct functions depending on the functioning and connectivity of the particular cortical area they involve (Haber, 2003; Yin and Knowlton, 2006). In this respect, the model presented here involves three BG-cortical macro loops that perform three different classes of selections relevant for solving the task:

- *Arm loop.* In the model, this sensorimotor loop performs the selection of actions involving the arm. In the brain, the selection of arm reaching actions involves the dorsal portions of the premotor cortex (PMC) and portions of the posterior parietal cortex (PC) and, in particular, the parietal reach region (PRR) (Wise et al., 1997; Luppino and Rizzolatti, 2000; Simon et al., 2002). These areas form re-entrant connections with the BG, in particular with the portion of striatum called putamen (Put) (Jaeger et al., 1993; Romanelli et al., 2005).

- *Oculomotor loop.* In the model, this loop performs the voluntary control of visual gaze (saccades). In the brain, the selection of eye movements involves the PC and, in particular, the lateral intraparietal cortex (LIP). The frontal eye fields (FEF) in the prefrontal cortex (PFC) are also a key cortical area for eye control. These areas form re-entrant loops with BG, in particular the portion of striatum called caudatum (Cau; Hikosaka et al., 2000).

- *Goal loop.* In the model, this loop encodes action outcomes and performs their selection

(goals). In the brain, the selection of goals strongly relies on PFC (Miller and Cohen, 2001) forming re-entrant connections with BG, in particular the nucleus accumbens portion of striatum (NAcc; Middleton and Strick, 2002; Haber, 2003).

Another key component of the model is the SC that generates phasic DA signals critical for learning. The SC of mammals is very sensitive to spatially localized changes in luminance caused, for example, by the appearance, disappearance, or movement of elements in the visual scene (Sparks, 1986; Wurtz and Albano, 1980). In such an event, the SC shows very fast sensory response (latency of about 40 ms; Jay and Sparks, 1987; Redgrave and Gurney, 2006) which activates the dopaminergic neurons of the substantia nigra pars compacta (SNc) and ventral tegmental area (VTA) and causes a phasic dopamine burst (Comoli et al., 2003; Dommett et al., 2005; May et al., 2009). The SC also shows a second response (latency around 200ms) which causes an orienting gaze shift to the region of space where the luminance change took place (Jay and Sparks, 1987; Sparks, 1986). FEF, and also LIP, project to SC, allowing the execution of voluntary eye movements (Hikosaka et al., 2000).

The phasic DA signal reaches, especially via SNc, cortical afferent terminals in striatum wherein it facilitates plasticity. Our hypothesis (Redgrave and Gurney, 2006) is that this plasticity could act to enhance the selection of the just-performed action in a transient way causing a *repetition bias* in the selection process for a transient period of time. This process has been demonstrated using biologically plausible models of spiking neurons (Gurney et al., 2009) and in an agent-based situation (Bolado-Gomez et al., 2009). Repetition bias allows representations of the action, its context, and its outcome to be repeatedly presented at neural circuits responsible for forming action-outcome associations thereby enhancing any Hebbian plasticity which might induce these associations, possibly involving DA (especially produced by VTA; Otani et al., 2003; Huang et al., 2004). Notice that according to our hypothesis, the animal brain devotes substantial resources (dopamine neurons, superior colliculus and cortico-striatal plasticity mechanisms) to bringing about this *autonomously generated*, transient change in policy.

An important aspect of the repetition bias is the fact that with repeated experience of the phasic stimuli the DA learning signal tends to be progressively diminished (Schultz, 1998). While the brain mechanisms that implement this cancellation are not clear, Redgrave et al. (2011) have hypothesised that this cancellation is the result of an active inhibition from another brain system. This might correspond to the BG (Smith et al., 2004), or the inhibitory inputs from other structures such as the lateral habenula (Balcita-Pedicino et al., 2011). The progressive attenuation of the phasic dopamine signal is critical for the transient nature of the repetition bias (Gurney et al., 2009; Bolado-Gomez et al., 2009) and has theoretical implications for optimal learning (Shah and Gurney, 2011). In this way, if the outcome is no more rewarding, the animal can disengage with the newly learned action and resume exploration that might lead to discover other interesting actions and outcomes to learn.

What are the brain mechanisms that allow the exploitation of actions once acquired? In the model, these are based on cortico-cortical connections as these have been shown to play an important role in the selection of actions based on current goals, and one view of such intra-cortical connectivity is that it implements an *internal model* of the action-outcome relationship (Gurney et al., 2012). In designing this aspect of the model, we took into consideration proposals about the macroscopic structure of the sensorimotor organisation of the brain Cisek and Kalaska (2010) (see Caligiore et al., 2010, for a model that captures the main features of this theory). These proposals start from evidence that the visuo-motor pathways in the brain are organised into two main neural pathways, the *dorsal* and *ventral* streams. (Mishkin and Ungerleider, 1982; Goodale and Milner, 1992). The key idea is that these two streams encode two main brain input-output mappings processing different aspects perception and action: the dorsal stream encodes multiple affordances

and actions in parallel, while the ventral stream, supported by the BG loops reviewed above, contributes to the selection of these affordances and actions. In particular, the dorsal stream contains the PC, which encodes affordances and implements the sensorimotor transformations that allow the animal to perform an on-line guidance of action execution. This stream also contains the PMC, which participates to the selection and preparation of actions (e.g., reach to a point in space, precision/power grasp, tear, etc.; Luppino and Rizzolatti, 2000; Caligiore et al., 2010). Importantly, the dorsal stream is organised in subsystems (Jeannerod, 1999; Rizzolatti and Matelli, 2003) that manage the motor control of different actuators, in particular the eyes (LIP in PC, and FEF in PFC; Snyder et al., 2000; Simon et al., 2002), the arms (PRR in PC, and PMCd; Wise et al., 1997; Simon et al., 2002), and the hands (anterior intraparietal area in PC, and inferior PMC; Rizzolatti et al., 2002).

The ventral stream involves the inferior temporal cortex (ITC), which plays an important role in object identification (Grill-Spector and Malach, 2004), and the PFC (Fuster, 2008). The PFC is a high-level multi-modal associative cortex receiving information from ITC about the resources available in the environment, and from limbic areas such as the amygdala (Amg, a subcortical system playing a key role in emotional processes, see below) about internal current needs and drives, and the consequent value assigned to resources (Wallis, 2007). On the basis of this information, the PFC forms goals and behavioural rules and uses them to exert a top-down bias on the selection of affordances and actions encoded within PC and the PMC in dorsal subsystems (Yeterian et al., 2011; Miller and Cohen, 2001; Caligiore et al., 2010). It is important to notice that all the neural connections considered here undergo learning processes, and these might be enhanced or made possible by neuromodulators such as DA (Otani et al., 2003; Huang et al., 2004).

Goal-directed behaviour is defined as a behaviour that is sensitive to the manipulations of the current *value* of the behavioural outcome (Balleine and Dickinson, 1998). The activation of goals is based on their current value encoded in Amg, a fundamental hub for the affective regulation of behaviour (Cardinal et al., 2002; Pitkänen et al., 1997; Balleine et al., 2003; Cardinal et al., 2002; Mirolli et al., 2010). Amg transmits information to PFC both directly and via the NAcc, a fundamental hub for goal-directed behaviour and DA-based energisation of behaviour (Voorn et al., 2004). In our model, the selection of goals is done by a hardwired mechanism mimicking the Amg goal selection at a phenomenological level. Our related work (Mannella et al., 2010) shows how Amg might be implemented in a more detailed fashion and support goal selection.

2.4 Computational and architectural description of the model

The detailed architecture of the model is illustrated in figure 3, and is based on the empirical evidence presented in Sec. 2.3. This figure, together with the model formal description below and the parameters indicated in tables 3, 4, 5, 6 in the Appendix, furnish all the information sufficient to replicate the model. Notice that, from one point of view, the model is rather simple; all neural units are described by similar equations and each component in the architecture, being grounded in biological data is not, therefore, 'novel' as such. Rather our aim was to select, quantitatively specify, and *integrate* a range of principles, mechanisms and ideas that, although not singularly novel, have hitherto, never been successfully combined into a functionally coherent whole. From another point of view then, the model is indeed complex, as the behaviour of our particular *combination* of neural components (with feedback loops and nonlinearities) will not necessarily be intuitive. In this respect, the most interesting properties of the model derive from the specific system-level connections shown in figure 3 and from the dynamical interplay between its components and the learning processes

described below.

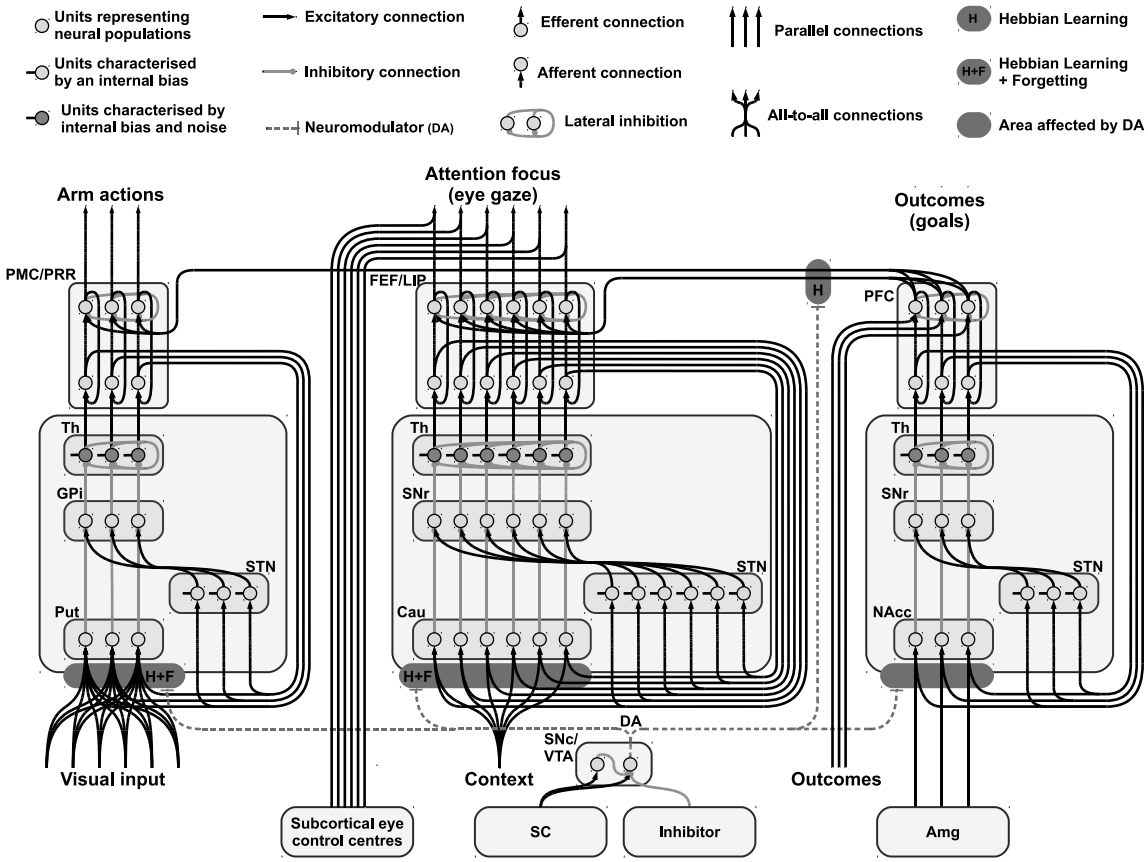


Figure 3: Detailed architecture of the model. Boxes implemented as rate-coded neural populations are shown with these populations as neural units (indicated by small circles). Boxes with text inside represent components whose function is implemented abstractly ('hardwired'). Other details are described in the figure legend and in the text. See table 2 in the Appendix for the anatomical acronyms.

The external inputs to the model have binary values (0 or 1). These inputs are: six inputs to the arm loop encoding the identity of the foveated object; one input to the oculomotor loop (always active at 1) representing the context; three inputs to the PFC representing the three possible changes of the environment (opening of the three boxes).

Each neural unit of the model intends to simulate the mean activity of a population of real neurons. This choice is justified by the type of study carried out here, in particular: (a) Our model is focussed on cognitive system-level phenomena, such as habitual and goal-directed behaviour, and slow learning processes, such as trial-and-error action acquisition and action-outcome encoding: these types of phenomena can be well captured with models based on firing rate units (Dayan and Abbott, 2001, pag. 229-231, Anastasio, 2010, pag. xvi-xvii); (b) The investigation of the target

phenomena does not require a fast response of neurons to rapidly changing inputs, one of the main limitations of firing rate units (Trappenberg, 2010, pag. 74–79). (c) The cognitive phenomena investigated here rely on the mean field response of whole populations of neurons, e.g. on the reciprocal influence of BG and cortex, rather than on the fast time scale reactions happening at the level of single neurons (Wilson and Cowan, 1972; Brunel and Wang, 2001; Bojak et al., 2003).

The basic building block of the model is thus a *leaky integrator unit* defined by a continuous-time differential equation as follows:

$$\tau_g \dot{u}_j = -u_j + J_j + b_j \tag{1}$$

where τ_g is a time constant, u_j is the activation potential of unit j, b_j is a baseline activation, and J_j represents the total net input to the unit. For all units apart from those in striatum, the net input is given by:

$$J_j = \sum_i w_{ji} y_i + I_j \tag{2}$$

where y_i is the output on unit i afferent to j (y_i depends on an activation function, see Eq. 4) and I_j is an external input (only present in the PFC outer layer). The connection weight between units i and j is denoted by w_{ji}. For striatal units the activation potential includes an external input and a dopaminergic modulation:

$$J_j = (\epsilon + \lambda d) \left(\sum_i w_{ji} y_i + I_j \right) \tag{3}$$

where d is the level of dopamine (see below) while ϵ and λ are two parameters that control respectively the strength of the whole input to a unit and the multiplicative effect of DA.

The activation of all units is defined using a positive saturation transfer function:

$$y_j = \left[\tanh \left(\alpha_g \left(u_j - \theta_g \right) \right) \right]^+ \tag{4}$$

where tanh(.) is the hyperbolic tangent function, α_g is a constant defining the slope of the hyperbolic function (related to the nucleus or group, g, of units to which j belongs), θ_g is a threshold parameter (per unit group), and $[.]^+$ is defined as $[x]^+ = 0$ if $x \leq 0$ and $[x]^+ = x$ if $x > 0$. Notice that $0 < y_j < 1$ for all j.

The output units of the model include the three output units of the arm loop (in the box labelled PMC/PRR in Figure 3, and representing the triggering of the three arm actions), and the six output units of the oculomotor loop (in the box labelled FEF/LIP in Figure 3, and representing foveation to the buttons and boxes). The output units have an activation and transfer function similar to the other units but they can trigger the execution of their corresponding action each time their activation overcomes a threshold of 0.8. A triggered action is maintained under execution only if the corresponding output unit remains active above such threshold, otherwise the execution of the action is aborted and the action fails to produce its effect.

The execution of an arm action lasted 1 sec, and the execution of saccade lasted 0.1 sec. If a box opened it stayed open for 2 sec. Time constants τ_g (Table 4 in the Appendix) were chosen to give behaviourally realistic times for actions and to allow the system to learn to successfully terminate triggered actions most of the times. The time step of the simulation update was 0.05 sec.

Within each of the three main BG-cortical loops, BG contained a direct, and hyperdirect pathway (the indirect pathway was not simulating as its regulatory effect was not needed). For each of the arm, oculomotor and goal loops, the direct pathway was constituted by striato-fugal pathways, Put→GPi, Cau→SNr, NAcc→SNr, respectively. The hyperdirect pathway involved STN and the

corresponding output nucleus (GPi, SNr, SNr). The BG component of each loop is similar to the model proposed by Leblois et al. (2006). This model is related to those of Gurney et al. (2001a,b) in their reliance on diffuse STN, and focussed striatal projections to form an off-centre, on-surround network for selection.

In addition, in the current model the thalamic complex (Th) is formed by a layer of units with lateral inhibitory connections. These connections implement a winner-take-all competition that gives an important contribution to the selection process working on the possible options available from the striatum. Intra-thalamic connectivity has also been proposed in this context by Humphries and Gurney (2002). In order for the model to implement trial-and-error reinforcement learning of new action combinations, there must be a source of variation or 'noise' in the selection of these actions. This is implemented in the thalamic grouping in which the activation u_j is subject to uniform noise in the interval $[-\nu, +\nu]$ at each time step.

The cortical component of each loop is formed by two reciprocally connected layers of units putatively corresponding to layers II/III (L2/3) – providing projections to other loops or external output – and layers IV/V (L4/5) – which project back to the BG. The units of L4/5 therefore encode the selected or 'winning' action channel which, via recurrence and competitive processing through the loops with BG, is able to reinforce the activation therein. This, in combination with the reciprocal excitatory connections within cortex, are parameterised such that the following two processe are able to operate. First, the selected channel continues to be selected after any input that initiated its selection has been removed. This 'lock-in' is needed so that the winning channel can remain active for the duration of its associated action, and thereby, any corresponding outcome. A second process is that selection of another channel must be possible when new salient input appears on that channel, and this must occur without any special 'reset' mechanism. These two processes are in tension but the model is able to let them interact appropriately.

In addition, the units of L2/3 have a high threshold and slope in their transfer function. Moreover, they have all-to-all lateral inhibitory connections that implement a winner-take-all competition (Figure 3). This endows the system with an extra selective function, aside the one within the Th, which allows a unique decision when there is a possible conflict between the actions 'suggested' by the top-down cortico-cortical connections from PFC and t he selections being fostered by the BG.

The model also contains four functionally abstract, non-neural ('hardwired') components: the SC, the DA inhibitor, the sub-cortical eye control centres, and Amg. The SC responds to luminance changes happening in the perceived scene: here it becomes active when any box opens. In this case SC activates the dopamine neuron units in SNc/VTA. The SNc/VTA is formed by two units representing excitatory and inhibitory sub-populations configured to produce an overall phasic response:

$$\tau_{SNc}\, \dot{u}_{in} = -u_{in} + y_{SC} \tag{5}$$

$$\tau_{SNc}\, \dot{u}_{ex} = -u_{ex} + [y_{SC} - u_{in}]^{+} \tag{6}$$

$$d^{*} = [\alpha_{SNc} \tanh(u_{ex} - \theta_{SNc})]^{+} \tag{7}$$

where u_{in} and u_{ex} are the activations of the inhibitory and excitatory sub-populations in SNc/VTA, y_{SC} is the output of SC, and d^{*} is the DA signal before the action of the 'inhibitor'.

The inhibitor subsumes, in a phenomenological fashion, the gradual attenuation of the phasic DA response if luminance changes happening in the same location in the environment are experienced several times, thereby becoming predictable/familiar. This function is implemented with three simple counters $N_k, k = 1, 2, 3$, of the openings of the three boxes so that the DA signal decreases

linearly in proportion to each of N_k, until it reaches zero for that box.

$$d = d^* - \mu N_k \tag{8}$$

where d is the dopamine signal, μ is a rate coefficient, and k is the box currently opening.

The model also has a hardwired component that drives the eye to saccade where luminance changes take place. This mechanism, which in animals is implemented by the SC and other subcortical systems controlling eye movements, overrides the voluntary actions selected by FEF/PC.

An final hardwired aspect of the model is the injection of activation into channels of the goal loop during the test phase. This activation, which mimics the attribution of values to the PFC outcomes by subcortical regions such as Amg, allows testing the capacity of the model to recall actions via the associated goals (see Mannella et al., 2010; Daw et al., 2005, for more biologically plausible models of this process).

The system undergoes two learning processes, the first involving the afferents to striatum in the arm and oculomotor loops, and the second involving the PFC→FEF/LIP and PFC→PMC/PRR cortico-cortical connections (see figure 3). DA also reaches the input connections of the goal loop, but here it only modulates the activation of the striatal units (according to equation 3) without giving place to any learning. This amounts to the assumption that the ability to associate value with a goal is already in place and does not require learning. The striatal learning process allows the arm loop to learn to associate suitable actions (e.g., press) to the seen objects (e.g., button 1) and it allows the oculomotor loop to learn to associate suitable saccades with the context unit so as to lead the system to focus on a particular portion of space. The learning process is based on a DA-dependent Hebbian learning rule:

$$y_j^+ = [y_j - \phi_{str}]^+ \tag{9}$$

$$d^+ = [d - \phi_d]^+ \tag{10}$$

$$\Delta w_{ji} = \eta_{str}\ d^+ y_j^+ \left(\hat{w}_{str} I_{ji} - w_{ji}\right) - \beta w_{ji} \tag{11}$$

where $\phi_{str} and \phi_d$ are thresholds for, respectively, the output and DA which have to be exceeded for learning to take place, as expressed in the quantities y_j^+ and d^+, w_{ji} is the input connection weight to the striatum, $\hat{w}_{str}$ is the maximum level that the weights can reach, and I_{ji} is the specific input i to the striatal unit j. The weights are also subject to an input-independent spontaneous decay with a constant rate β. The core component of the rule, $y_j^+ (\hat{w}_{str} I_{ji} - w_{ji})$, contains a Hebbian learning term, $y_j^+(\hat{w}_{str} I_{ji})$, and a term $-y_j^+ w_{ji}$ that leads each weight to decay in proportion to its value and the output y_j^+ (Willshaw and von der Malsburg, 1976, and see Rolls and Treves, 1998 page 72 for a discussion). Overall, the component implies that each weight w_{ji} progressively moves towards $\hat{w} I_{ji}$ in proportion to y_j^+. The input-independent spontaneous decay of the rule implies that when DA is suppressed by the inhibitor ($d = 0$) the weights progressively approach zero. This is at the basis of the transient nature of the focussing of oculomotor and arm actions caused by the repetition bias.

The learning process involving the cortico-cortical connections drives the association of each outcome encoded in the PFC (e.g., 'box 1 opens') with a particular location in space selected by oculomotor loop (e.g., 'look at button 1') and a particular action of the arm loop (e.g., 'press'). The learning process is also based on a Hebbian learning rule, but invokes DA-dependent *eligibility traces* g_j of the arm and oculomotor loop units in PMC/PRR, and FEF/LIP respectively. The eligibility trace is needed as the outcome representation in PFC occurs some time after the actions that caused

it – the saccade to the button and the button press – and it is *these* actions that need associating with the outcome, rather than the subsequent saccade to the box. In this respect, DA caused by the opening of the box occurs immediately after the activation of the cortical units triggering the saccade (e.g., 'look button 1') and the arm action (e.g., 'press') and is coincident with non-zero outputs in these units. It can therefore be used to prime an eligibility trace process which can subsequently be used in a Hebb-like rule involving the PFC representation of the outcome (opening of the box) when this eventually occurs. DA is indeed caused by the arm action if it succeeds in opening a box, and such action follows the saccade and takes sometime to be executed. The eligibility trace is charged when units of L2/3 within the oculomotor and the arm loops activate, and this event is immediately followed by a DA signal. Thus:

$$\tau_{tr}\dot{g}_j = -g_j + \zeta y_j d \tag{12}$$

$$\Delta w_{ji} = \eta_{ctx}\ g_j y_i (\hat{w}_{ctx} - w_{ji}) \tag{13}$$

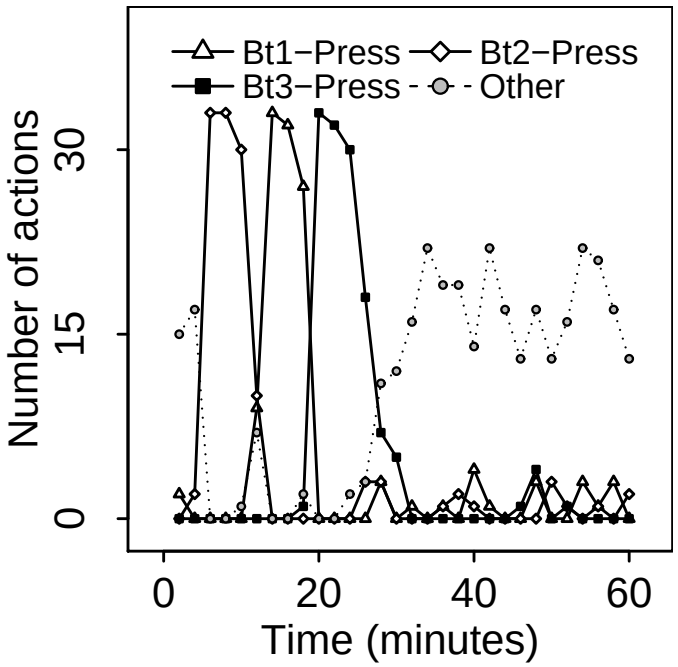
where τ_{tr} is a time constant, ζ is a rate coefficient, w_{ji} is the connection weight between the PFC unit i and the FEF/LIP or PMC/PRR unit j, η_{ctx} is a learning coefficient, a_i is the unit of PFC, and $\hat{w}_{ctx}$ is a maximum value reachable by w_{ji}.

The parameters of the model were set by manual search to obtain a stable, functioning system, whose *behaviour* was qualitatively similar to the behaviour of real subjects in the board experiment. Indeed, it was not appropriate, given the system-level nature of the model, to try to set the model parameters on the basis of *physiological* data as the links between the two are not known.

3 Results

This section first shows the behaviour of the full, 'intact' model (also referred to as the 'base' model) shown in in Figure 3, and the evolution of its trained connection weights during the learning phase. Then, to show the role of key elements of the model, the section compares the performance of the intact model with the performance of versions of the model where such elements are lesioned. Four lesions were performed: lesion of the input connection weights to the Put, or to the Cau, or to both, and lesion of the dopamine inhibitor. The aim of the lesions of Put and Cau was to investigate the (differential) effects that a reduced arm-action or attentional focusing would have had on learning. The lesion of the inhibitor was, instead, directed at investigating the role played in learning by the key feature of IMs related to their transient nature. Note that, currently, there are no empirical data to verify the lesion results, so these have to be considered predictions of the model testable in future experiments.

Figure 4 shows the actions performed by the model during a learning phase of 60 mins; here time is divided into 30 time bins lasting 2 min each. The figure shows a specific example of training because it is not possible to plot an average of the behaviours of different simulation runs. The reasons is that systems from different simulations focus on exploring the buttons in different order or with asynchronous focusing periods. However, all simulation runs showed qualitatively similar behaviours. Initially (first 2 bins), the model performs a random exploration of the environment (executing random saccades and arm actions), but soon focusses to look at, and press, button 2. However, this focussing is transient: it lasts for about 8 min, after which the model focusses on button 1 for about 10 min and then on button 3 for about 8 min. After these focussed activities, the system again engages in a random exploration of the environment. The focussing of actions in this way is exactly what we mean by 'repetition bias'.



.

Figure 4: Example of behaviour of the intact model during the training phase. Other repetitions of the experiment produce qualitatively similar results. The y-axis shows the number of executions of action compounds of the type 'look at button x, press button x' (labelled *Bt1-Press*, *Bt2-Press*, and *Bt3-Press* for the three buttons), and also the number of executions of all other action compounds considered together (*Other*). Data are reported for time bins lasting two mins each.

We now illustrate the system behaviour and internal functioning during learning. Figure 5 shows the dynamics of the striatal input connection weights of the arm and oculomotor loops, and also the cortico-cortical connection weights from the PFC to FEF/LIP and PMC/PRR, during training. Figure 5a,b shows how the striatal input connections of both the arm and oculomotor loops undergo an initial increase in strength followed by a decrease: this transient change keeps the activities of the system focussed on each button for about 5-10 min. Each transient focus is driven by the initial high level of DA produced by the sight of the opening of a box followed by a progressive inhibition of it due to the inhibitor update. Figure 5c,d shows how this transient focussing leads to quickly develop the cortico-cortical connection weights encoding the action-outcome associations linking the various outcomes with the eye and arm actions that led to cause them. In contrast to the transient striatal weights, these weights permanently store the 'knowledge' about the task (action-outcome contingencies). Overall these results show that the model is indeed capable of acquiring an effective goal-directed behaviour on the basis of the DA learning signal caused by IMs and the transient learning processes of the striatum (repetition bias).

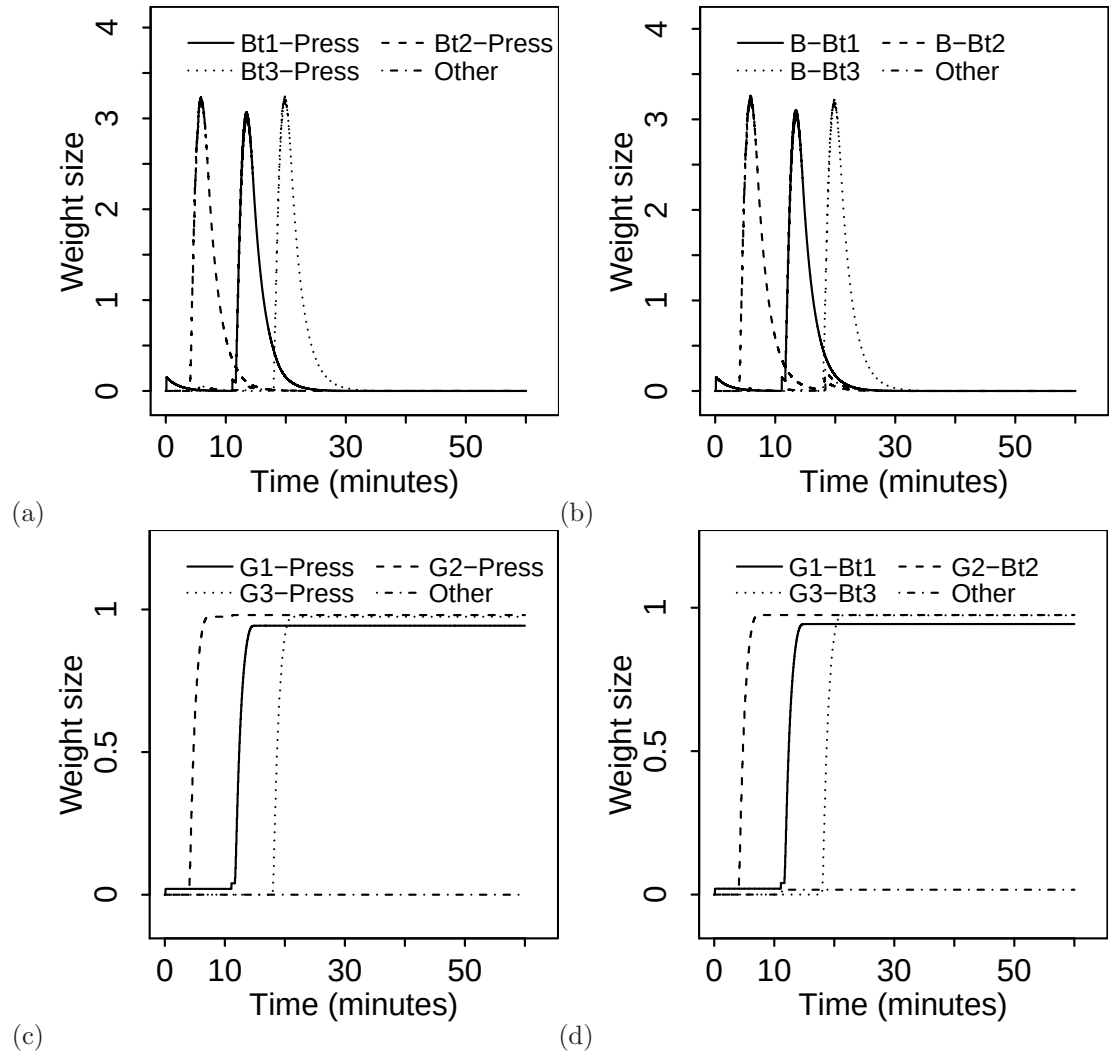


Figure 5: Example of development of the trained connection weights of the model during the learning phase. (a) Input connection weights to Put (arm loop striatum). *Bt1-Press, Bt2-Press, Bt3-Press*: connection weights between the input units representing button 1, 2 and 3, and the BG channel of the press action. *Other*: average of all other input connection weights. (b) Input connection weights to Cau (oculomotor loop striatum). *B-Bt1, B-Bt2, B-Bt3*: connection weights between the context unit and the BG channels related to looking at button 1, 2, and 3. (c) Cortico-cortical connection weights from the PFC (goal loop cortex) to the PMC/PRR (arm loop cortex). *G1-Press, G2-Press, G3-Press*: connection weights between the PFC units encoding the three goals and the PMC/PRR unit encoding the press action. (d) Cortico-cortical connection weights from the PFC to the FEF/LIP (oculomotor loop cortex). *G1-Bt1, G2-Bt2, G3-Bt3*: connection weights between the PFC units encoding the goals and the FEF/LIP units related to looking at button 1, 2, and 3.

An evaluation based on the test procedure was performed at 6 minute intervals during the learning phase to evaluate performance as learning progressed. Each evaluation comprised 50 repetitions of the test phase in which the three goals were sequentially activated, each one for 2 mins (single time bin), making a test phase of 6 mins duration in all. We introduce a simple notation to explain clearly how the performance of the model was measured in this evaluation. Let m_{rg} be the number of correct box-openings for goal $g = 1, 2, ..., 3$ during the repetition $r = 1, 2, ..., 50$ of the test-phase, and let $M_j = \frac{\sum_g m_{rg}}{3}$. Let σ_r be the standard deviation of the m_{rg} for repetition r. Performance was measured in two ways: in terms of the mean number of correct box-openings $\langle M_r \rangle$ measured over the 50 repetitions; and in terms of the mean of the standard deviations $\langle \sigma_r \rangle$ again computed over the 50 repetitions. The first metric gives an information on how good the system is in accomplishing the goals, given the amount of learning time available in the training phase. The second metric gives an information on how differentiated the performance is for the different goals, again given the amount of learning time available.

The performance based on these two metrics is shown in Figures 6 and 7 respectively. The intact model performs the correct actions (fixation of the correct button and press) very efficiently after the whole training. The performance decreases progressively with the decrease of the learning time available. This shows that the exploration and action acquisition driven by IMs during the learning phase led the system to acquired the needed cortical connections weights ('internal models') that later allow it to recall suitable actions to pursue the desired goals.

The input connections to the arm and oculomotor loop striatum were lesioned in isolation, or together, to quantify the effects of the repetition bias of the arm or oculomotor action son the speed of learning. Note that when the striatal input connections to either one of the loops are lesioned, such loop selects actions randomly. This still allows the formation of cortico-cortical connections, but slows down their development. The figures show that the intact model has the highest rate of action acquisition, followed by the two conditions of Put or Cau lesions, the condition of simultaneous Put and Cau lesions, and finally the inhibitor lesion.

The lesion of the Cau in the oculomotor loop slows the learning process as the system wastes time looking at (and hence interacting with) the boxes rather than the buttons. The lesion of the Put in the arm loop slows learning as the system performs many of the inconsequential 'dummy' actions on the buttons. Figure 6 shows that the Put lesion leads to a lower performance than the Cau lesion for intermediate training times, as the former model still has Cau intact and so tends to focus on one button before passing to another one. As the intact model takes almost the entire training phase to progress through all buttons, at intermediate learning times the model with lesioned Put (but intact Cau) terminates learning while having still little knowledge of some buttons (those on which it did not have time to focus learning). In contrast, the Cau lesion (no attentional focus) still allows the system to learn to interact with all buttons to a certain extent with any duration of learning, so giving it a better overall performance with respect to the Put lesioned model when the available learning time is short.

This tension between an even spread of learning across the buttons versus a focussed, button-by-button approach is manifest in figure 7. When the Cau input is present (Intact and Put conditions), at intermediate times the variance in performance across the three goals is large, as good performance is skewed preferentially to only one or two goals. In contrast, when the Cau input is lesioned (Cau and Put+Cau conditions) the system tends to uniformly learn all action-outcomes associations at the same time, giving a smaller variance at all stages. The lesion of both Put and Cau further slows learning as the system wastes time to both look at, and interact with, the boxes (due

to the Cau lesion) or to perform the wrong actions on buttons (due to the Put lesion).

Figures 6 and 7 also show the effect of a lesion of the inhibitor on learning. In this condition the capacity to pursue the goals remains quite low (figure 6). The reason is that the DA produced by a box opening does not decrease, so the system remains obsessively focussed on looking at, and performing the press actions, on one button only. This implies that, at the end of learning, the system is capable of pursuing only one goal, as shown by the high variance of performance across different goals independently of the duration of the learning phase (figure 7).

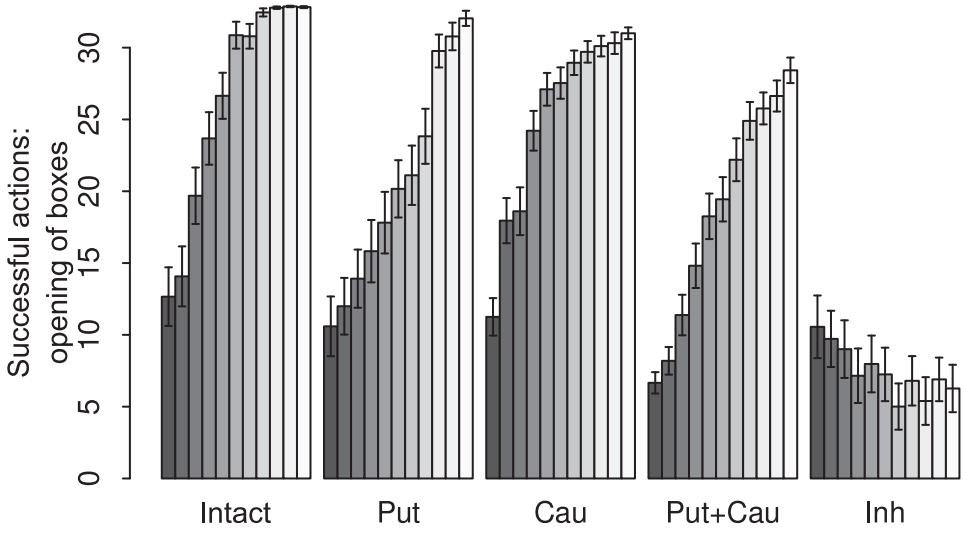


Figure 6: Performance of the non-lesioned model ('Intact') and four versions of the model where lesions were performed to the input connections to the arm loop (*Put*), oculomotor loop (*Cau*), both loops (*Put-Cau*), or to the inhibitor (*Inh*). For each condition, each individual histogram bar is defined by the the metric $\langle M_g \rangle$ (mean number of correct box openings per goal over 50 tests, see text). Performance was measured at 6 minute intervals, corresponding to the histogram bars, over the entire learning phase (so the last histogram bar of each condition reports the performance of the corresponding model after a full learning period of 60 mins). The histogram bars also report the standard error over the 50 repetitions.

The explanations given above, in terms of relative focus or spread of action selection, are confirmed in Figure 8 which shows the behaviour of a single instance of the model in the four lesion conditions, allowing comparison with the behaviour of the intact model in figure 4. In particular, Figure 8a confirms that the lesion of the Put still allows the model to focus its interaction on the buttons one by one. In comparison to the intact model, however, the interaction with each button is longer (about 15-20 min) as the actions performed with the arm are random. This, in turn, causes a less frequent production of the DA learning signal, a less frequent update of the inhibitor, and, as a consequence, the slowing down of the learning of the striatum and cortex.

In the case of the Cau lesion, Figure 8b shows that the system interacts with all buttons at

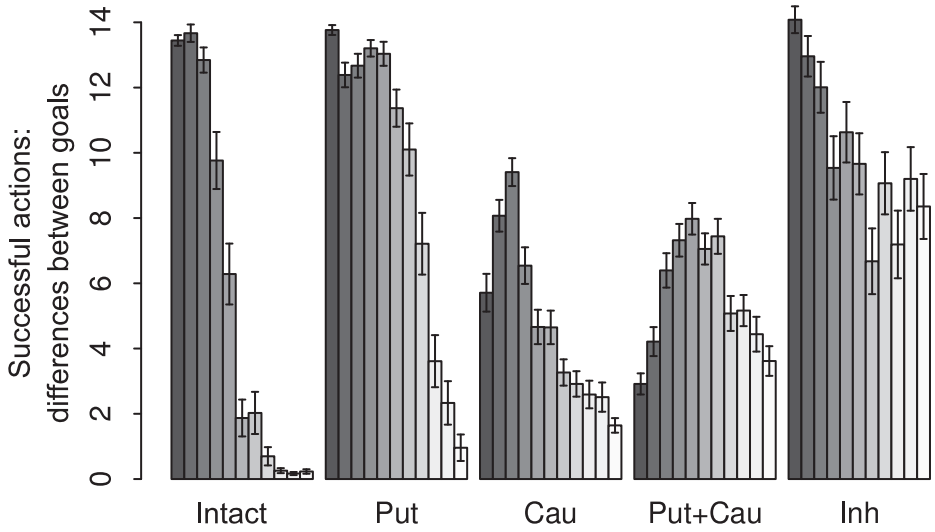


Figure 7: Data related to the same experiments described in figure 6 but defined by the metric $\langle \sigma_j^M \rangle$ (based on the standard deviation of box-openings for the different goals, averaged over the 50 repetitions of the test, see text). This is done at 6 min intervals over the entire learning phase (histogram bars). Histogram bars also report the standard error over the 50 repetitions.

the same time. This eventually leads the system to learn to act suitably on them. However, we conjecture that this behaviour would not scale up well with the number of actions to learn (only three in this case) and that 'spreading' of learning over large numbers of actions (as encountered in a real ecological setting, for example) will eventually lead to failure to learn any of them. When both Put and Cau are lesioned (Figure 8c) the rate of foveation to closed boxes and the performance of irrelevant actions is quite high and this makes learning inefficient. Finally, when the inhibitor is lesioned (Figure 8d), the system focusses attention and action on only one button and does not disengage from it even after a prolonged training, so impeding the acquisition of skills and knowledge related to other buttons and boxes.

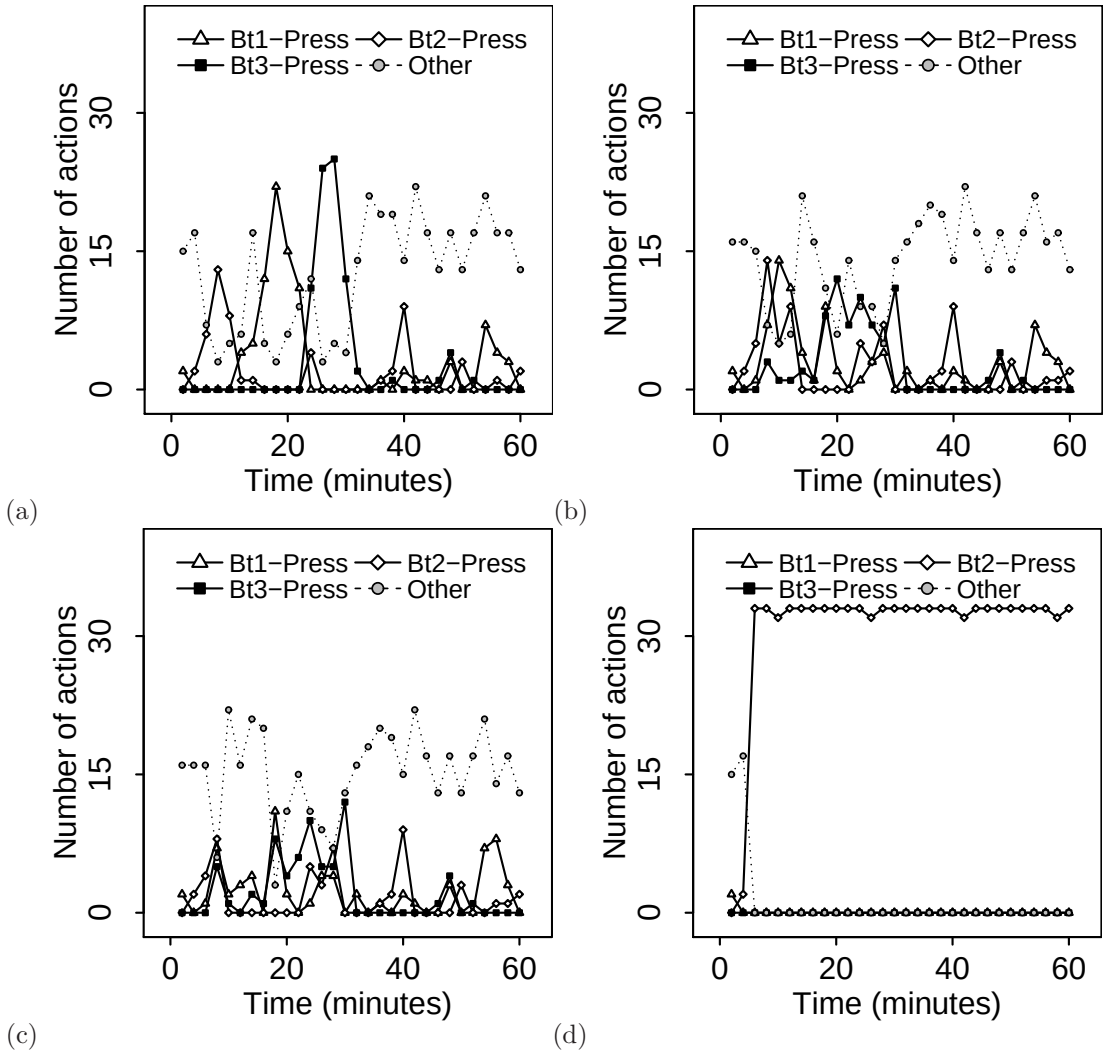
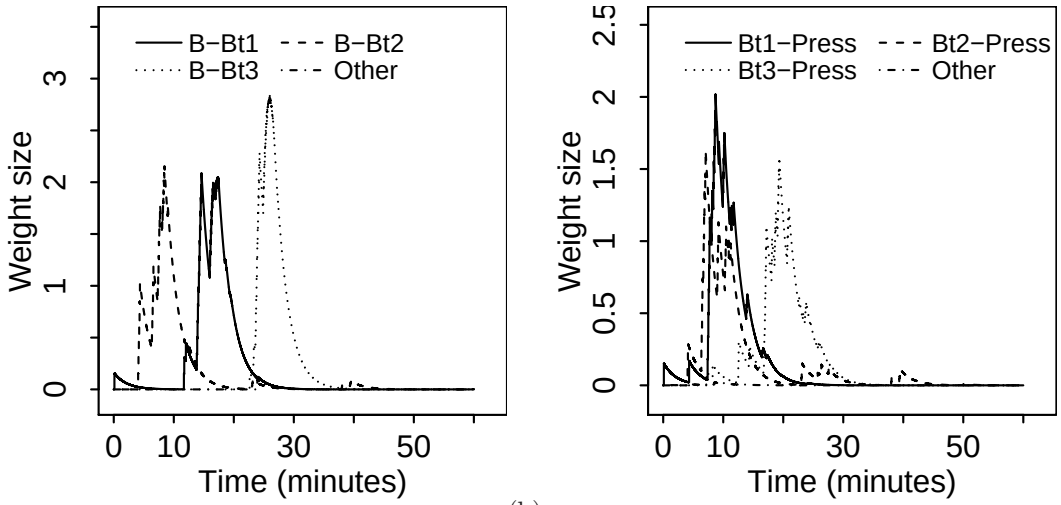


Figure 8: Example of actions performed by the model affected by different types of lesions, plotted as in figure 4. Other repetitions of the experiments produce qualitatively similar results. (a) Lesion of the input connections to Put (arm loop). (b) Lesion of the input connections to Cau (oculomotor loop). (c) Lesion of the input connections to both Put and Cau. (d) Lesion of the inhibitor.

Figure 9 shows the effects of lesions of the Put or the Cau on the development of the input connection weights of the Cau and Put, respectively. The dynamics of these weights underlie the behaviour of the respective systems illustrated above. Figure 9a indicates that in the case of the Put lesion, the system learns the Cau weights related to the different buttons one by one, further confirming that the learning processes proceed in sub-phases driven by the focussing of attention. In the case of the Cau lesion, instead, figure 9b shows that the acquisition of knowledge related to the three buttons tends to take place in parallel.



(a) (b)

Figure 9: Example of the effects that the lesion of the Put (arm loop) or the Cau (oculomotor loop) cause on the development of the input connection weights to the Cau and Put, respectively. (a) Put lesion, Cau connetion weights. (b) Cau lesion, Put connection weights.

Figure 10 shows the development of the cortico-cortical connection weights with the four lesions. These weights represent the action-outcome knowledge (internal models) produced by the training processes illustrated above. In the case of the Put lesion (arm loop; figure 10a), the system acquires each action-outcome association relatively fast as it still focuses the learning processes on the different buttons one by one. However, the process is slightly slower than in the case of the intact model (see figure 5d). In the case of the Cau lesion (oculomotor loop; figure 10b), the system acquires the knowledge on the three action-outcome associations in parallel as it does not focus on single experiences. The same happens if the Put and Cau are lesioned together, even if learning now proceed at an even slower pace (figure 10c). In the case the inhibitor is lesioned (figure 10d) the system learns only the action-outcome association useful to open the second box as it remains focussed on the second button for the whole duration of the learning phase.

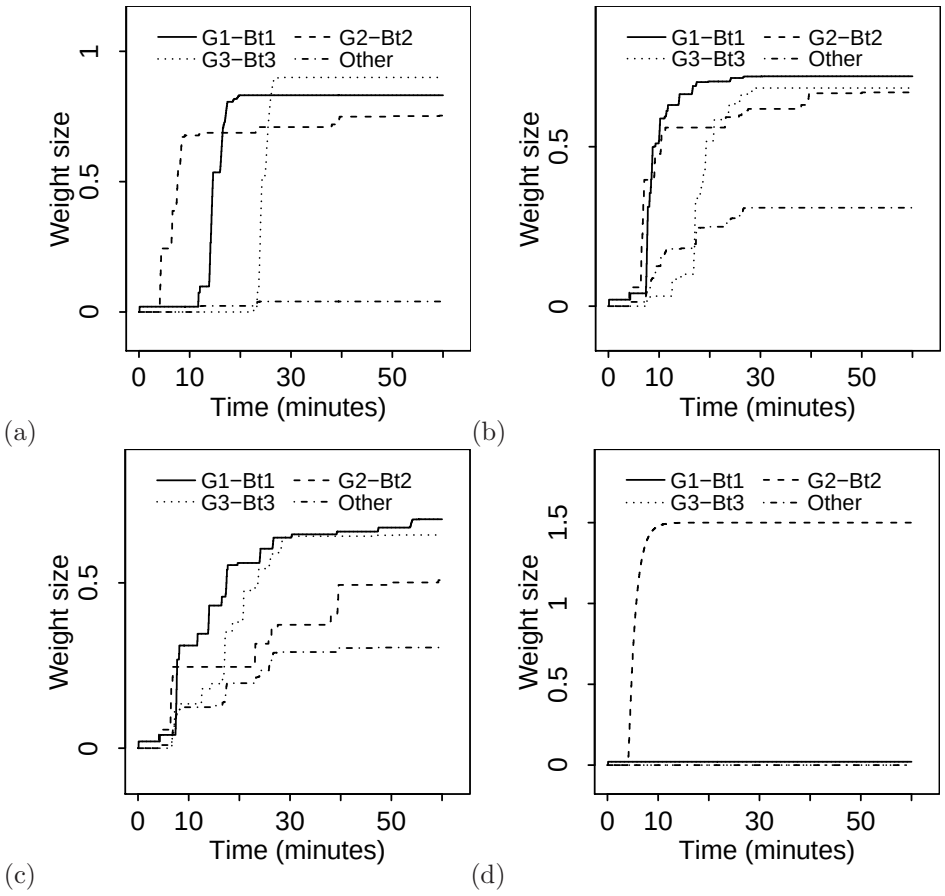


Figure 10: Example of dynamics of the connection weights linking PFC to FEF/LIP that the model affected by the four lesions exhibits during the learning phase (the connection weights from PFC to PMC/PRR, not reported here for brevity, exhibit similar dynamics). Data are plotted as done in figure 5d. (a) Lesion of the input connections to Put (arm loop). (b) Lesion of the input connections to Cau (oculomotor loop). (c) Lesion of the input connections to both Put and Cau. (d) Lesion of the inhibitor.

4 Discussion and Conclusions

This research has proposed a novel model that furnishes an operational hypothesis on how intrinsic motivations can support the acquisition of a repertoire of actions and the encoding of the related action-outcome associations, and on how these associations can be later used to recall the actions when these might be useful for adaptation (i.e., to accomplish an extrinsic reward). The value of the model resides in the fact that, based on an architecture constrained at the system-level with relevant neuroscientific evidence, it *integrates* a number of important mechanisms and processes important for IM: (a) an overall sensorimotor architecture, based on striato-cortical loops, that includes an oculomotor and an arm-control circuit (allowing the agent to explore the environment both perceptually and operantly), and a goal circuit supporting goal-directed behaviour; (b) a mechanism for guiding learning based on novel events, based on SC and a progressive inhibition of DA signals; (c) a repetition bias mechanism that supports an effective focussing of the learning resources on different available experiences; (d) the learning of internal models of action-outcome contingencies, and their later exploitation to recall actions via a value-based reactivation of goals within PFC. Several of these mechanisms and processes were proposed in Redgrave and Gurney (2006) based on neuroscientific evidence and theoretical analyses. In this paper, however, we have specified these mechanisms quantitatively and added additional hypotheses relating to the mechanisms and neural substrate for the learning of action-outcomes, the recall of goals and their effect on action selection, the synchronisation of bottom-up and top-down processes, the management of the emergent dynamics of the learning processes based on the repetition bias. In so doing, the model represents the first relatively complete implementation of the original theory.

The main results are that (i) a simple, but biologically inspired implementation of repetition bias supports efficient learning of action outcome associations. (ii) These associations may be formed across disparate action components (i.e. in both the oculomotor and arm-control circuits). (iii) The different roles of repetition bias on visual action focussing and focussing of arm movements is revealed by a series of selective lesions; this leads to predictions that may be tested in animals. (iv) These lesions also highlight the role of repetition bias in focussing action acquisition sequentially for, when this does not occur, learning of the internal models in cortex is compromised. The spreading of 'action focus' that ensues from poor repetition bias does not prevent this learning in the toy domain used here, where there are few action combinations, but the combinatorial explosion of these combinations in a more realistic setting, may prevent adequate association learning altogether. (v) The model shows how the internal models of action-outcome contingency learned under IM, can be recruited for goal-directed activity.

Preliminary data from the empirical experiments run with children are broadly consistent with the model's behaviour. The model has also produced various predictions based on lesions that might be tested in future experiments. Preliminary results with a version of the model embodied in a humanoid robot are encouraging and will be reported elsewhere.

4.1 Biological and Psychological Issues

Various elements of the model open interesting issues that need to be further investigated or developed in future work. A first element is the generation of the dopamine learning signal by the SC. The generation of learning signals by intrinsic motivations, for example based on DA, is a critical aspect for the model. The current version of the model focusses on the generation of a learning signal by the SC when a change of luminance happens in the environment. This has been shown

to play an important role in driving striatal plasticity underlying trial-and-error learning (see Redgrave and Gurney, 2006, for a review). An open problem on this is how far intrinsically motivated acquisition of behaviours can be supported by such a mechanism, considering that the SC cannot distinguish between different textures, colours, shapes, etc. In this respect, a more sophisticated capability of detecting the consequences of actions seems needed to acquire some skills (for instance, learning to arrange two familiar objects in a particular novel spatial relation, e.g. for learning to stick one toy block on top of another). Other brain components might generate learning signals in these situations. For example the hippocampus has been shown to respond to novel objects, or novel spatial or temporal combinations of familiar objects, and to activate dopaminergic neurons on this basis (Lisman and Grace, 2005). A related, and currently debated issue, is the nature of the DA signal itself: is this related to (extrinsic) rewards (Schultz, 1998), or phasic stimuli (neutral or rewarding) as in this model (Redgrave and Gurney, 2006), or both? Future investigations, both empirical and theoretical, are needed to fully disantangle this issue (e.g., see Santucci et al., 2010, for a theoretical proposal and Mirolli et al., subm, for a model that aim to reconcile the two views).

A second issue concerns the learning rules, especially in striatum, where the dynamics of the weights are complex and have to show a phasic behaviour with 'spontaneous decay'. These rules are currently somewhat phenomenological and, while suitable for a high level treatment of the kind considered in this model, there is plenty of scope for incorporating more biological realism. Data from studies in cortico-striatal plasticity have provided a complex and often confusing picture, because of the dependence of plasticity on D1 and D2 type dopamine receptors, and on dopamine levels themselves. Recently a study by Shen et al. (2008) using powerful transgenic, *in vitro* techniques, has shed light on these issues, and shows a rich variety of cortico-striatal plasticity under spike-timing-dependent-plasticity (STDP) protocols. We have recently developed a modelling framework to explain this data and, remarkably, the *in vitro* data are entirely consistent with the action-outcome learning schema (Gurney et al., 2009). Related, biologically plausible learning rules have also been investigated in a study of repetition bias in behaving agents (Bolado-Gomez et al., 2009). It remains a challenge to integrate these new rules into the kind of system-level model described here but this promises more natural accounts of the dynamics of cortico-striatal plasticity.

Another issue opened up by the model concerns the timing of the learning processes taking place within the striatum, the cortex, and the (hardwired) inhibitor. This timing is important as the release from repetition bias ('unfocussing' of the action selection) due to the weight decay in striatum (in turn driven by the dopamine decrease caused by the inhibitor) has to have the same time scale as the learning of action-outcomes of the cortex. Indeed, if the unlearning of the striatum is too fast, there might not be enough time for cortex to learn action-outcomes, whereas if it is too slow it might lead the system to waste time on activities whose effects have already been learned. In relation to this point, the biological literature indicates that intra-cortical learning leading to the automatisation of behaviour is usually thought to take longer than the striatal learning processes (cf. Ashby et al., 2010). However, the learning speed of cortex might increase in cases where it is innervated by DA (Otani et al., 2003; Huang et al., 2004). Another solution to the problem might also come from forgetting processes involving the inhibitor: an attenuation of the effectiveness of this component might allow the system to periodically re-establish 'interest' in previous experience, thereby allowing a refinement of the knowledge on action-outcomes acquired by cortex. Many experiments on biological learning show that, when learning is repeated in sessions held on different days, both the behavioural skills and the synaptic plasticity tend to increase during each session, but also to partially regress towards initial levels from one session to the other (Lieberman, 1993).

Another important element of the model that needs to be further investigated is the dopamine

'inhibitor' (currently hardwired). The function of this component is to drive the progressive attenuation of the dopaminergic learning signal. As discussed in Sec. 2.3, various possibilities exist that could support such a mechanism (Redgrave et al., 2011). These include the inhibitory efferents of the SNr, which also project to the SC (Hikosaka et al., 2000), or the BG efferents to dopaminergic neurons; another possibility might be the inhibitory projections of lateral habenula to dopaminergic areas (Balcita-Pedicino et al., 2011). This issue is further discussed below from a computational perspective.

There are also some open issues related to the capacity of the model to recall actions once learned. A first issue is related to the internal mechanisms which drive recall of outcomes (goals), currently hardwired in the model. Internal drives based on EMs are a primary source of the recall of actions when the organism needs to satisfy particular needs: this is indeed a key function of 'motivations' (Panksepp, 1998). The goal-directed literature recalled in the introduction proposes that sub-cortical structures play a key role in this processes, in particular the Amg (Cardinal et al., 2002; Mirolli et al., 2010). Amg has been shown to play a key role in the assignment of biological value to goals and hence in their recall, an important process at the basis of 'goal-directed behaviour' (Balleine and Dickinson, 1998, Balleine et al., 2003), especially via NAcc (Pennartz et al., inpr). The model presented here will be updated in the future with the addition of an Amg component based on the model presented by Mannella et al. (2010). This module will be capable of learning to assign a value to goals (e.g., by seeing a food item in a box the model will 'value' the goal of opening that box) and on this basis to recall the skills previously acquired with IMs.

4.2 Computational Issues

The introduction distinguished between knowledge-based IMs (KB-IMs) and competence-based IMs (CB-IMs) relying on the typology proposed in Oudeyer and Kaplan (2007). Further, it specified that, as argued by Mirolli and Baldassarre (inpr), there can be KB-IM and CB-IM *mechanisms* can be both used for the acquisition of either knowledge or competence (so we can talk of KB-IM and CB-IM *functions*). This raises the question on the nature of the algorithm presented here, based on the SC and the inhibitor, with respect to these classes. The answer is that such an algorithm is a KB-IM mechanism serving the function of competence acquisition. It is a KB-IM mechanism as it measures the novelty of salient events (box openings) on the basis of how frequently they have been experienced but independently of the capacity of the system to cause them (competence). The function served by this mechanism, however, is the acquisition of competence, namely the acquisition of action-outcome contingencies.

The particular type of KB-IM mechanism used here has been investigated within the computational literature by Schmidhuber (1991b) (see Schmidhuber, 2010 for a review of this and other similar approaches). The model proposed in this work is formed by a predictor component (that learns to predict the next sensations on the basis of the current state and the planned actions), and a reinforcement learning component (equivalent to the dopamine-based trial-and-error learning used in the model proposed here). The reinforcement learning component learns to perform actions on the basis of a reward signal given by the (absolute value) of the *error of prediction* of the predictor. The interaction of these two components lead the system to perform actions that bring the agent in states of the world that cause a high error of the predictor. This allows the predictor to learn the new situations and so to drive the agent to seek novel experiences.

Interestingly, Schmidhuber (1991a, 1999) (see Schmidhuber, 2010, for a review) has later proposed that this mechanism might have difficulties if the predictor cannot learn to predict the

consequences of some actions, in particular because the world is stochastic or because of the computational limitations of the predictor. The author suggests that a solution to this problem is to use the *improvement of the prediction error* instead of the prediction error as a reinforcement signal. What is interesting is that from a biological perspective it seems that the SC works on the basis of a prediction error and so it should be affected by the problem highlighted by Schmidhuber. However, the problem might indeed be solved by other additional mechanisms. For example, Santucci et al. (2010) and Mirolli et al. (subm) propose a possible computational solution to the problem: if a hierarchical organisation of actions is used, and learning resources are allocated to different regions of sensorimotor space in proportion to the learning rate of *skills* in those regions, as done in Schembri et al. (2007b,a,c) (see Baldassarre and Mirolli, 2012, for a review), then the problem raised by Schmidhuber could be solved. Indeed, in this case, the focussing of learning resources on different regions would depend on the actual skills acquired (competence), not on the capacity of the predictor to predict, used to train the skills themselves (notice, however, that in this hypothesis the progressive inhibition of the signal used to train the skills would not be required).

The problem of coupling between the learning speed of the striatum and the cortex has also implications from a computational perspective. This is also related to the issue of using a KB-IM mechanism, as done here, for the acquisition of competence. The use of KB-IM mechanisms to acquire competence is quite common in the computational literature, since the classic models on IM (Schmidhuber, 1991b,a; Oudeyer et al., 2007) (but there are important models that use CB-IM mechanisms to acquire competence, e.g., Singh S. and Chentanez, 2005; Hart and Grupen, 2011; Schembri et al., 2007c). However, as highlighted in Mirolli and Baldassarre (inpr), this approach might have some limitations if the goal of the system is mainly to acquire competence due to the fact that the process leading to acquire knowledge might be faster or slower than the process leading to acquire competence. For example, in the model considered here, a too fast learning of the inhibitor would not allow a full acquisition of action-outcomes contingencies related to one button. On the other side, a too slow learning of the inhibitor would lead to waste time to explore a button when the competence related to it has already been acquired. A solution to this problem might be based on an inhibitor that incorporates a mechanism that automatically ensures a coupling between the two learning processes. This result can be achieved with an inhibitor that measures the actual acquired competence instead of the acquired knowledge. For example, in the case of our model, one might aim to create an inhibitor that learns to actively inhibit phasic DA based on the actual success of the system in achieving the outcome, with a complete cessation of phasic DA only when the skill/action-outcome associations are fully learned. A successful coupling between two learning processes is for example achieved by the reinforcement learning actor-critic model (Sutton and Barto, 1998). In this model, the learning of the critic component (a predictor of reward) closely follows the learning of the actor component (which acquires competence). This idea is indeed exploited in the model of Baldassarre and Mirolli, 2012 mentioned above to build a CB-IM mechanism. One might investigate if and how it is possible to exploit a similar mechanism to implement the inhibitor within the current model, also considering that the actor-critic model is often used to capture the reward prediction error signalled by DA and the acquisition of instrumental skills by BG (Houk et al., 1995; Joel et al., 2002).

From a computational perspective, another direction in which the model should be developed concerns the representation and use of goals. First, goals should be encoded and learned on the basis of actual experiences from the environment (this aspect is rather abstract in the current model). Second, they might be activated in an anticipatory fashion not only during action recall but also during learning, giving rise to a 'goal-based learning' process for which the activation of goals might

aid learning in several ways, for example by focussing experience on relevant portions of space and by generating learning signals (Baldassarre, 2002, 2003).

Although we think all these open issues call for further developments and refinements of the model in the future, the model architecture represents a novel framework to further develop the theoretical understanding and empirical investigation on how actions and actions-outcomes are first learned on the basis of IMs and then exploited based on goals activated by EMs.

Acknowledgements

This research has received funds from the 7th Framework Programme of the European Community (FP7/2007-2013), *Challenge 2 - Cognitive Systems, Interaction, Robotics*, Grant Agreement No. ICT-IP-231722, Project *IM-CLeVeR - Intrinsically Motivated Cumulative Learning Versatile Robots*.

References

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

Anastasio, T. J. (2010). *Tutorial on neural systems modelling*. Sunderland, MA: Sinauer Associated.

Ashby, F. G., B. O. Turner, and J. C. Horvitz (2010). Cortical and basal ganglia contributions to habit learning and automaticity. *Trends Cogn Sci 14*(5), 208–215.

Balcita-Pedicino, J. J., N. Omelchenko, R. Bell, and S. R. Sesack (2011). The inhibitory influence of the lateral habenula on midbrain dopamine cells: ultrastructural evidence for indirect mediation via the rostromedial mesopontine tegmental nucleus. *J Comp Neurol 519*(6), 1143–1164.

Baldassarre, G. (2002). *Planning with neural networks and reinforcement learning*. Phd thesis, Computer Science Department, University of Essex, Colchester, UK.

Baldassarre, G. (2003). Forward and bidirectional planning based on reinforcement learning and neural networks in a simulated robot. In M. Butz, O. Sigaud, and P. Grard (Eds.), *Adaptive behaviour in anticipatory learning systems*, Volume 2684 of *Lecture Notes in Artificial Intelligence*, pp. 179–200. Berlin: Springer Verlag.

Baldassarre, G. (2011). What are intrinsic motivations? a biological perspective. In A. Cangelosi, J. Triesch, I. Fasel, K. Rohlfing, F. Nori, P.-Y. Oudeyer, M. Schlesinger, and Y. Nagai (Eds.), *Proceedings of the International Conference on Development and Learning and Epigenetic Robotics (ICDL-EpiRob-2011)*, pp. E1–8. New York: IEEE.

Baldassarre, G. and M. Mirolli (2012). Deciding which skill to learn when: Temporal-difference competence-based intrinsic motivation (TD-CB-IM), a mechanism that uses the td-error as intrinsic reinforcement in hierarchical systems. In G. Baldassarre and M. Mirolli (Eds.), *Intrinsically Motivated Learning in Natural and Artificial Systems*. Berlin: Springer-Verlag. (this volume).

Balleine, B. W. and A. Dickinson (1998). Goal-directed instrumental action: contingency and incentive learning and their cortical substrates. *Neuropharmacology 37*(4-5), 407–419.

Balleine, B. W., S. A. Killcross, and A. Dickinson (2003). The effect of lesions of the basolateral amygdala on instrumental conditioning. *J Neurosci 23*(2), 666–675.

Berlyne, D. E. (1966). Curiosity and exploration. *Science 143*, 25–33.

Berridge, K. C., J. W. Aldridge, K. R. Houchard, and X. Zhuang (2005). Sequential super-stereotypy of an instinctive fixed action pattern in hyper-dopaminergic mutant mice: a model of obsessive compulsive disorder and tourette's. *BMC Biol 3*, 4.

Berridge, K. C. and T. E. Robinson (1998). What is the role of dopamine in reward: hedonic impact, reward learning, or incentive salience? *Brain Res Brain Res Rev 28*(3), 309–369.

Bojak, I., T. Oostendorp, A. Reid, and R. Kotter (2003). Connecting mean field models of neural activity to eeg and fmri data. *Brain Topography 23*, 139–149.

Bolado-Gomez, R., J. Chambers, and K. Gurney (2009). The basal ganglia and the 3-factor learning rule: reinforcement learning during operant conditioning. In J. Triesch, J. Lcke, G. Pipa, C. Rothkopf, and J. Zhu (Eds.), *Frontiers in Computational Neuroscience. Conference Abstracts: Bernstein Conference on Computational Neuroscience*, pp. 154–155.

Brown, J., D. Bullock, and S. Grossberg (1999). How the basal ganglia use parallel excitatory and inhibitory learning pathways to selectively respond to unexpected rewarding cues. *J Neurosci 19*(23), 10502.

Brunel, N. and X. Wang (2001). Effects of neuromodulation in a cortical network model of object working memory dominated by recurrent inhibition. *Journal of computational neuroscience 11*(1), 63–85.

Calabresi, P., B. Picconi, A. Tozzi, and M. D. Filippo (2007). Dopamine-mediated regulation of corticostriatal synaptic plasticity. *Trends Neurosci 30*(5), 211–219.

Caligiore, D., A. Borghi, D. Parisi, and G. Baldassarre (2010). TRoPICALS: A computational embodied neuroscience model of compatibility effects. *Psychol Rev 117*, 1188–1228.

Cardinal, R. N., J. A. Parkinson, J. Hall, and B. J. Everitt (2002). Emotion and motivation: the role of the amygdala, ventral striatum, and prefrontal cortex. *Neurosci Biobehav Rev 26*(3), 321–352.

Carelli, R. M., M. Wolske, and M. O. West (1997). Loss of lever press-related firing of rat striatal forelimb neurons after repeated sessions in a lever pressing task. *J Neurosci 17*(5), 1804–1814.

Chevalier, G. and J. M. Deniau (1990). Disinhibition as a basic process in the expression of striatal functions. *Trends Neurosci 13*(7), 277–280.

Cisek, P. and J. F. Kalaska (2010). Neural mechanisms for interacting with a world full of action choices. *Annu Rev Neurosci 33*, 269–298.

Comoli, E., V. Coizet, J. Boyes, J. P. Bolam, N. S. Canteras, R. H. Quirk, P. G. Overton, and P. Redgrave (2003). A direct projection from superior colliculus to substantia nigra for detecting salient visual events. *Nat Neurosci 6*(9), 974–980.

Cope, A. and K. N. Gurney (2011). A biologically based model of active vision. In *Proceedings of AISB'11 - Architectures for Active Vision*, York, UK.

Crabtree, J. W. and J. T. R. Isaac (2002). New intrathalamic pathways allowing modality-related and cross-modality switching in the dorsal thalamus. *J Neurosci 22*(19), 8754–8761.

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

Dayan, P. and L. Abbott (2001). *Theoretical neuroscience: Computational and mathematical modeling of neural systems*. Boston, MA: MIT Press.

Dommett, E., V. Coizet, C. D. Blaha, J. Martindale, V. Lefebvre, N. Walton, J. E. W. Mayhew, P. G. Overton, and P. Redgrave (2005). How visual stimuli activate dopaminergic neurons at short latency. *Science 307*(5714), 1476–1479.

Doya, K. (1999). What are the computations of the cerebellum, the basal ganglia and the cerebral cortex? *Neural Netw 12*(7-8), 961–974.

Fuster, J. M. (2008). *The prefrontal cortex* (fourth ed.). Oxford: Elsevier.

Goodale, M. A. and A. D. Milner (1992). Separate visual pathways for perception and action. *Trends Neurosci 15*(1), 20–25.

Grill-Spector, K. and R. Malach (2004). The human visual cortex. *Annual Review of Neuroscience 27*, 649–677.

Gurney, K., M. Humphries, and P. Redgrave (2009). Cortico-striatal plasticity for action-outcome learning using spike timing dependent eligibility. In *BMC Neuroscience*, Volume 10, pp. E135.

Gurney, K., N. Lepora, A. Shah, A. Koene, and P. Redgrave (2012). Action discovery and intrinsic motivation: a biologically constrained formalisation. In G. Baldassarre and M. Mirolli (Eds.), *Intrinsically Motivated Learning in Natural and Artificial Systems*. Berlin: Springer-Verlag. (this volume).

Gurney, K., T. Prescott, and P. Redgrave (2001a). A computational model of action selection in the basal ganglia I: A new functional anatomy. *Biol Cybern 84*, 401–410.

Gurney, K., T. Prescott, and P. Redgrave (2001b). A computational model of action selection in the basal ganglia II: analysis and simulation of behaviour. *Biol Cybern 84*, 411–423.

Gurney, K., T. Prescott, J. Wickens, and P. Redgrave (2004). Computational models of the basal ganglia: from robots to membranes. *Trends Neurosci 27*(8), 453–459.

Gurney, K. N. (2009). Reverse engineering the vertebrate brain: Methodological principles for a biologically grounded programme of cognitive modelling. *Cognitive Computation 1*(1), 29–41.

Haber, S. N. (2003). The primate basal ganglia: parallel and integrative networks. *J Chem Neuroanat 26*, 317330.

Harlow, H. F. (1950). Learning and satiation of response in intrinsically motivated complex puzzle performance by monkeys. *Journal of Comparative and Physiological Psychology 43*, 289–294.

Hart, S. and R. Grupen (2011). Learning generalizable control programs. *IEEE Transactions on Autonomous Mental Development 3*(1), 216–231.

Hikosaka, O. (1998). Neural systems for control of voluntary action–a hypothesis. *Adv Biophys 35*, 81–102.

Hikosaka, O., Y. Takikawa, and R. Kawagoe (2000). Role of the basal ganglia in the control of purposive saccadic eye movements. *Physiol Rev 80*(3), 953–978.

Houk, J. C., J. L. Adams, and G. B. Andrew (1995). A model of how the basal ganglia generate and use neural signals ghat predict reinforcement. In J. C. Houk, J. L. Davids, and D. G. Beiser (Eds.), *Models of Information Processing in the Basal Ganglia*, pp. 249–270. Cambridge, MA: The MIT Press.

Houk, J. C., J. L. Davids, and D. G. Beiser (Eds.) (1995). *Models of Information Processing in the Basal Ganglia*. Cambridge, MA: The MIT Press.

Huang, Y.-Y., E. Simpson, C. Kellendonk, and E. R. Kandel (2004). Genetic evidence for the bidirectional modulation of synaptic plasticity in the prefrontal cortex by d1 receptors. *Proc Natl Acad Sci U S A 101*(9), 3236–3241.

Hull, C. L. (1943). *Principles of Behavior*. New York, NY: Appleton-century-crofts.

Humphries, M. and K. Gurney (2002). The role of intra-thalamic and thalamocortical circuits in action selection. *Network: Computation in Neural Systems 13*, 131–156.

Jaeger, D., S. Gilman, and J. W. Aldridge (1993). Primate basal ganglia activity in a precued reaching task: preparation for movement. *Exp Brain Res 95*(1), 51–64.

Jay, M. F. and D. L. Sparks (1987). Sensorimotor integration in the primate superior colliculus. i. motor convergence. *J Neurophysiol 57*(1), 22–34.

Jeannerod, M. (1999). Visuomotor channels: Their integration in goal-directed prehension. *Human Movement Science 18*(2-3), 201–218.

Joel, D., Y. Niv, and E. Ruppin (2002). Actor-critic models of the basal ganglia: new anatomical and computational perspectives. *Neural Netw 15*(4-6), 535–547.

Kakade, S. and P. Dayan (2002). Dopamine: generalization and bonuses. *Neural Netw 15*(4-6), 549–559.

Kandel, E. R., J. H. Schwartz, and T. M. Jessell (2000). *Principles of Neural Science* (fourth ed.). New York, USA: McGraw–Hill.

Kish, G. (1955). Learning when the onset of illumination is used as the reinforcing stimulus. *Journal of Comparative and Physiological Psychology 48*(4), 261–264.

Kumaran, D. and E. A. Maguire (2007). Which computational mechanisms operate in the hippocampus during novelty detection? *Hippocampus 17*(9), 735–748.

Leblois, A., T. Boraud, W. Meissner, H. Bergman, and D. Hansel (2006). Competition between feedback loops underlies normal and pathological dynamics in the basal ganglia. *J Neurosci 26*(13), 3567–3583.

Lieberman, D. A. (1993). *Learning: behavior and cognition* (second ed.). Pacific Grove, CA: Brooks/Cole Publishing Company.

Lisman, J. E. and A. A. Grace (2005). The hippocampal-vta loop: controlling the entry of information into long-term memory. *Neuron 46*(5), 703–713.

Luppino, G. and G. Rizzolatti (2000). The Organization of the Frontal Motor Cortex. *News Physiol Sci 15*, 219–224.

Mannella, F., M. Mirolli, and G. Baldassarre (2010). The interplay of pavlovian and instrumental processes in devaluation experiments: a computational embodied neuroscience model tested with a simulated rat. In C. Tosh and G. Ruxton (Eds.), *Modelling Perception With Artificial Neural Networks*. Cambridge: Cambridge University Press.

May, P. J., J. G. McHaffie, T. R. Stanford, H. Jiang, M. G. Costello, V. Coizet, L. M. Hayes, S. N. Haber, and P. Redgrave (2009). Tectonigral projections in the primate: a pathway for pre-attentive sensory input to midbrain dopaminergic neurons. *Eur J Neurosci 29*(3), 575–587.

Middleton, F. A. and P. L. Strick (2002). Basal-ganglia 'projections' to the prefrontal cortex of the primate. *Cereb Cortex 12*(9), 926–935.

Miller, E. K. and J. D. Cohen (2001). An integrative theory of prefrontal cortex function. *Annu Rev Neurosci 24*, 167–202.

Mirolli, M. and G. Baldassarre (inpr). Functions and mechanisms of intrinsic motivations. In G. Baldassarre and M. Mirolli (Eds.), *Intrinsically motivated learning in natural and artificial systems*. Berlin: Springer-Verlag.

Mirolli, M., G. Baldassarre, and V. G. Santucci (subm). Phasic dopamine as a prediction error of intrinsic and extrinsic reinforcement driving both action acquisition and reward maximization: A simulated robotic study. *Neural Net*.

Mirolli, M., F. Mannella, and G. Baldassarre (2010). The roles of the amygdala in the affective regulation of body, brain, and behaviour. *Connection Science 22*(3), 215–245.

Mishkin, M. and L. G. Ungerleider (1982). Contribution of striate inputs to the visuospatial functions of parieto-preoccipital cortex in monkeys. *Behav Brain Res 6*(1), 57–77.

Nambu, A., H. Tokuno, and M. Takada (2002). Functional significance of the cortico-subthalamopallidal 'hyperdirect' pathway. *Neurosci Res 43*(2), 111–117.

Otani, S., H. Daniel, M.-P. Roisin, and F. Crepel (2003). Dopaminergic modulation of long-term synaptic plasticity in rat prefrontal neurons. *Cereb Cortex 13*(11), 1251–1256.

Oudeyer, P., F. Kaplan, and V. Hafner (2007). Intrinsic motivation systems for autonomous mental development. *IEEE Transactions on Evolutionary Computation 11*(2), 265–286.

Oudeyer, P.-Y. and F. Kaplan (2007). What is intrinsic motivation? A typology of computational approaches. *Frontiers in Neurorobotics 1*, 6.

Panksepp, J. (1998). *Affective Neuroscience: The Foundations of Human and Animal Emotions*. Oxford: Oxford Unversity Press.

Pennartz, C., R. Ito, P. Verschure, F. Battaglia, and T. Robbins (inpr). The hippocampalstriatal axis in learning, prediction and goal-directed behavior. *Cell*.

Pitkänen, A., V. Savander, and J. E. LeDoux (1997). Organization of intra-amygdaloid circuitries in the rat: an emerging framework for understanding functions of the amygdala. *Trends Neurosci 20*(11), 517–523.

Redgrave, P. and K. Gurney (2006). The short-latency dopamine signal: a role in discovering novel actions? *Nature Rev Neurosci 7*(12), 967–975.

Redgrave, P., T. J. Prescott, and K. Gurney (1999). The basal ganglia: a vertebrate solution to the selection problem? *Neuroscience 89*(4), 1009–1023.

Redgrave, P., N. Vautrelle, and J. N. J. Reynolds (2011). Functional properties of the basal ganglia's re-entrant loop architecture: selection and reinforcement. *Neuroscience 198*, 138–151.

Reynolds, J. N. J. and J. R. Wickens (2002). Dopamine-dependent plasticity of corticostriatal synapses. *Neural Net 15*(4-6), 507–521. PMID: 12371508.

Rizzolatti, G., L. Fogassi, and V. Gallese (2002). Motor and cognitive functions of the ventral premotor cortex. *Curr Opin Neurobiol 12*(2), 149–154.

Rizzolatti, G. and M. Matelli (2003). Two different streams form the dorsal visual system: anatomy and functions. *Exp Brain Res 153*(2), 146–157.

Rolls, E. T. and A. Treves (1998). *Neural networks and brain function*. Oxford: Oxford Unversity Press.

Romanelli, P., V. Esposito, D. W. Schaal, and G. Heit (2005). Somatotopy in the basal ganglia: experimental and clinical evidence for segregated sensorimotor channels. *Brain research. Brain research reviews 48*, 112–28.

Santucci, V. G., G. Baldassarre, and M. Mirolli (2010). Biological cumulative learning through intrinsic motivations: a simulated robotic study on development of visually-guided reaching. In B. Johansson, E. Sahin, and C. Balkenius (Eds.), *Proceedings of the Tenth International Conference on Epigenetic Robotics (EpiRob2010)*, pp. 121–128. Lund, Sweden: Lund University.

Sara, S. J. (2009). The locus coeruleus and noradrenergic modulation of cognition. *Nature Rev Neurosci 10*(3), 211–223.

Sara, S. J., A. Vankov, and A. Herv (1994). Locus coeruleus-evoked responses in behaving rats: a clue to the role of noradrenaline in memory. *Brain Res Bull 35*(5-6), 457–465.

Schembri, M., M. Mirolli, and G. Baldassarre (2007a). Evolution and learning in an intrinsically motivated reinforcement learning robot. In L. M. Almeida e Costa Fernando, Rocha, E. Costa, I. Harvey, and A. Coutinho (Eds.), *Advances in Artificial Life. Proceedings of the 9th European Conference on Artificial Life (ECAL2007)*, Volume 4648 of *Lecture Notes in Artificial Intelligence*, pp. 294–333. Berlin: Springer Verlag.

Schembri, M., M. Mirolli, and G. Baldassarre (2007b). Evolving childhood's length and learning parameters in an intrinsically motivated reinforcement learning robot. In L. Berthouze, P. G. Dhristiopher, M. Littman, H. Kozima, and C. Balkenius (Eds.), *Proceedings of the Seventh International Conference on Epigenetic Robotics*, Volume 134, pp. 141–148. Lund: Lund University.

Schembri, M., M. Mirolli, and G. Baldassarre (2007c). Evolving internal reinforcers for an intrinsically motivated reinforcement-learning robot. In Y. Demiris, D. Mareschal, B. Scassellati, and J. Weng (Eds.), *Proceedings of the 6th International Conference on Development and Learning*, Piscataway, NJ, pp. E1–6. IEEE.

Schmidhuber, J. (1991a). Curious model-building control systems. In *Proceedings of the International Joint Conference on Neural Networks*, Volume 2, pp. 1458–1463.

Schmidhuber, J. (1991b). A possibility for implementing curiosity and boredom in model-building neural controllers. In J. A. Meyer and S. W. Wilson (Eds.), *Proceedings of the International Conference on Simulation of Adaptive Behavior: From Animals to Animats*, Cambridge, MA, pp. 222–227. MIT Press/Bradford Books.

Schmidhuber, J. (1999). Artificial curiosity based on discovering novel algorithmic predictability through coevolution. In *Evolutionary Computation, 1999. CEC 99. Proceedings of the 1999 Congress on*, Volume 3.

Schmidhuber, J. (2010). Formal theory of creativity, fun, and intrinsic motivation (1990–2010). *IEEE Transactions on Autonomous Mental Development 2*(3), 230–247.

Schultz, W. (1998). Predictive reward signal of dopamine neurons. *J Neurophysiol 80*, 127.

Schultz, W. (2010). Dopamine signals for reward value and risk: basic and recent data. *Behav Brain Funct 6*(1), 24.

Shah, A. and K. Gurney (2011). Dopamine-mediated action discovery promotes optimal behavior for free. *BMC Neuroscience 12*(Suppl 1), P138.

Shen, W., M. Flajolet, P. Greengard, and D. J. Surmeier (2008). Dichotomous dopaminergic control of striatal synaptic plasticity. *Science (New York, N.Y.) 321*(5890), 848–851.

Shepherd, G. and S. Grillner (2010). *Handbook of brain microcircuits*. Oxford: Oxford University Press.

Simon, O., J. F. Mangin, L. Cohen, D. L. Bihan, and S. Dehaene (2002). Topographical layout of hand, eye, calculation, and language-related areas in the human parietal lobe. *Neuron 33*(3), 475–487.

Singh, S., R. Lewis, A. Barto, and J. Sorg (2010). Intrinsically motivated reinforcement learning: An evolutionary perspective. *IEEE Transactions on Autonomous Mental Development 2*(2), 70–82.

Singh S., B. A. G. and N. Chentanez (2005). Intrinsically motivated reinforcement learning. In L. K. Saul, Y. Weiss, and L. Bottou (Eds.), *Advances in Neural Information Processing Systems 17: Proceedings of the 2004 Conference*, Cambridge, MA. The MIT Press.

Skinner, B. (1938). *The Behavior of Organisms*. New York, NY: Appleton-Century-Crofts.

Smith, Y., D. V. Raju, J.-F. Pare, and M. Sidibe (2004). The thalamostriatal system: a highly specific network of the basal ganglia circuitry. *Trends Neurosci 27*(9), 520–527.

Snyder, L. H., A. P. Batista, and R. A. Andersen (2000). Intention-related activity in the posterior parietal cortex: a review. *Vision Res 40*(10-12), 1433–1441.

Sparks, D. L. (1986). Translation of sensory signals into commands for control of saccadic eye movements: role of primate superior colliculus. *Physiol Rev 66*(1), 118–171.

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

Taffoni, F., D. Formica, M. Schiavone, G. andScorcia, A. Tomasetti, E. Polizzi di Sorrentino, G. Sabbatini, V. Truppa, F. Mannella, V. Fiore, M. Mirolli, M. Mirolli, G. Baldassarre, , E. Visalberghi, F. Keller, and E. Guglielmelli (inpr). The mechatronic board: A tool to study intrinsic motivations in humans, monkeys, and humanoid robots. In G. Baldassarre and M. Mirolli (Eds.), *Intrinsically motivated learning in natural and artificial systems*. Berlin: Springer Verlag.

Taffoni, F., M. Vespignani, D. Formica, G. Cavallo, E. Polizzi di Sorrentino, G. Sabbatini, V. Truppa, E. Visalberghi, M. Mirolli, G. Baldassarre, F. Keller, and E. Guglielmelli (2012). A mechatronic platform for behavioural analysis of non human primates. *Journal of Integrative Neuroscience 11*(1), 87–101.

Trappenberg, T. P. (2010). *Fundamentals of computational neuroscience* (2 ed.). Oxford: Oxford University Press.

von Hofsten, C. (1982). Eye-hand coordination in newborns. *Dev Psychol 18*(3), 450–461.

Voorn, P., L. J. M. J. Vanderschuren, H. J. Groenewegen, T. W. Robbins, and C. M. A. Pennartz (2004). Putting a spin on the dorsal-ventral divide of the striatum. *Trends Neurosci 27*(8), 468–474.

Wallis, J. D. (2007). Orbitofrontal cortex and its contribution to decision-making. *Annu Rev Neurosci 30*, 31–56.

White, R. W. (1959). Motivation reconsidered: The concept of competence. *Psychol Rev 66*, 297–333.

Wickens, J. R. (2009). Synaptic plasticity in the basal ganglia. *Behav Brain Res 199*(1), 119–128.

Wickens, J. R., J. N. J. Reynolds, and B. I. Hyland (2003). Neural mechanisms of reward-related motor learning. *Curr Opin Neurobiol 13*(6), 685–690.

Willshaw, D. J. and C. von der Malsburg (1976). How patterned neural connections can be set up by self-organization. *Proc R Soc Lond B Biol Sci 194*(1117), 431–445.

Wilson, H. and J. Cowan (1972). Excitatory and inhibitory interactions in localized populations of model neurons. *Biophysical journal 12*(1), 1–24.

Wise, S. P., D. Boussaoud, P. B. Johnson, and R. Caminiti (1997). Premotor and parietal cortex: corticocortical connectivity and combinatorial computations. *Annu Rev Neurosci 20*, 25–42.

Wurtz, R. H. and J. E. Albano (1980). Visual-motor function of the primate superior colliculus. *Annu Rev Neurosci 3*, 189–226.

Yeterian, E. H., D. N. Pandya, F. Tomaiuolo, and M. Petrides (2011). The cortical connectivity of the prefrontal cortex in the monkey brain. *Cortex 48*(1), 58–81.

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

Yu, A. and P. Dayan (2003). Expected and unexpected uncertainty: ACh and NE in the neocortex. In *Advances in Neural Information Processing Systems 15: Proceedings of the 2002 Conference*, pp. 157–164. Cambridge, MA: The MIT Press.

Appendix

Table 2: Acronyms used in the paper.

Brain components	
Amygdala	Amg
Basal ganglia	BG
Caudatum	Cau
Frontal eye fields	FEF
Globus pallidus internum	GPi
Inferior temporal cortex	ITC
Lateral intraparietal cortex	LIP
Layer II/III of cortex	L2/3
Layer IV/V of cortex	L4/5
Nucleus accumbens	NAcc
Parietal cortex	PC
Parietal reach region	PRR
Prefrontal cortex	PFC
Premotor cortex	PMC
Putamen	Put
Striatum	Str
Superior colliculus	SC
Substantia nigra pars compacta	SNc
Substantia nigra pars reticulata	SNr
Subthalamic nucleus	STN
Ventral tegmental area	VTA
Thalamus	Th

Other	
Dopamine	DA
Extrinsic motivations	EMs
Intrinsic motivations	IMs
Knowledge-based IMs	KB-IMs
Competence-based IMs	CB-IMs
Long term depression	LTD
Long term potentiation	LTP

Table 3: Connection weights within each striato-cortical loop.

	Loop		
Inter-layer connections	**Put**	**Cau**	**NAcc**
Input → Str	+0.4	+0.4	+0.4
L4/5 → Str	+1.0	+1.0	+0.5
L4/5 → STN	+1.6	+1.0	+1.0
STN → GPi/SNr	+1.4	+0.8	+3.4
Str → GPi/SNr	−3.0	−3.0	−3.0
GPi/SNr → Th	−2.0	−2.0	−2.0
Thal → L4/5	+2.8	+2.8	+2.8
L4/5 → L2/3	+1.0	+1.0	+1.0
L2/3 → L4/5	+0.5	+0.5	+0.5
Intra-layer connections	**Put**	**Cau**	**NAcc**
Th self-connection	+1.2	+1.2	+0.3
Th lateral-connection	−8.0	−4.0	−1.0
L2/3 lateral-connection	−2.0	−2.0	−2.0
Neuromodulatory connections	**Put**	**Cau**	**NAcc**
DA-dependent Str	+4.0	+4.0	+4.0
DA-independent Str	+0.2	+0.2	+0.2

Table 4: Parameters regulating the activation of units within the striato-cortical loops.

		Loop	
Decays (τ_g)	**Put**	**Cau**	**NAcc**
Str	0.3	0.3	0.3
STN	0.3	0.3	0.3
GPi/SNr	0.3	0.3	0.3
Th	0.3	0.3	0.3
L4/5	1.2	0.3	1.2
L2/3	0.3	0.3	0.3
Baseline potentials (b_g)	**Put**	**Cau**	**NAcc**
Str	0.0	0.0	0.0
STN	0.5	0.5	0.5
GPi/SNr	0.0	0.0	0.0
Th	2.0	2.0	2.0
L4/5	0.0	0.0	0.0
L2/3	0.0	0.0	0.0
Output thresholds (θ_g)	**Put**	**Cau**	**NAcc**
Str	0.0	0.0	0.0
STN	0.0	0.0	0.0
GPi/SNr	0.0	0.0	0.0
Th	0.0	0.0	0.0
L4/5	0.6	0.6	0.6
L2/3	0.8	0.8	0.8
Output slopes (α_g)	**Put**	**Cau**	**NAcc**
Str	1.0	1.0	1.0
STN	1.0	1.0	1.0
GPi/SNr	1.0	1.0	1.0
Th	1.0	1.0	1.0
L4/5	1.0	1.0	1.0
L2/3	20.0	20.0	20.0
Noise range ($[-\nu, +\nu]$)			
Th	$[-3.5, +3.5]$	$[-3.5, +3.5]$	$[-3.5, +3.5]$

Table 5: Parameters regulating the activation of units within the SNc/VTA.

Parameters	**Values**
Decay (τ_{SNc})	0.1
Output thresholds (θ_{SNc})	0.0
Output slopes (α_{SNc})	1.0

Table 6: Parameters that regulate the learning processes.

Inhibitor	
Novelty decay (μ)	0.001
Learning rates	
Input→Put (η_{str})	0.06
Input→Cau (η_{str})	0.06
PFC→FEF/LIP (η_{ctx})	0.001
PFC→PMC/PRR (η_{ctx})	0.001
Deacy rates	
Input→Put (β)	0.001
Input→Cau (β)	0.001
Saturation thresholds	
Input→Put ($\hat{w}_{str}$)	50
Input→Cau ($\hat{w}_{str}$)	50
PFC→FEF/LIP ($\hat{w}_{ctx}$)	1.5
PFC→PMC/PRR ($\hat{w}_{ctx}$)	1.5
Learning thresholds	
DA (ϕ_d)	0.6
Str (ϕ_{str})	0.95
Cortical traces	
Decay (τ_{tr})	8
Charging coefficient (ζ)	60