Experimental predictions drawn from a computational model of sign-trackers and goal-trackers

Florian Lesaint[a,b,*], Olivier Sigaud[a,b], Jeremy J. Clark[c], Shelly B. Flagel[d,e,f], Mehdi Khamassi[a,b]

[a] *Sorbonne Universités, UPMC Univ Paris 06, UMR 7222, Institut des Systèmes Intelligents et de Robotique, F-75005, Paris, France*
[b] *CNRS, UMR 7222, Institut des Systèmes Intelligents et de Robotique, F-75005, Paris, France*
[c] *Department of Psychiatry and Behavioral Sciences, University of Washington, Washington, USA*
[d] *Department of Psychiatry, University of Michigan, Ann Arbor, Michigan, USA*
[e] *Molecular and Behavioral Neuroscience Institute, University of Michigan, Ann Arbor, Michigan, USA*
[f] *Department of Psychology, University of Michigan, Ann Arbor, Michigan, USA*

Abstract

Gaining a better understanding of the biological mechanisms underlying the individual variation observed in response to rewards and reward cues could help to identify and treat individuals more prone to disorders of impulsive control, such as addiction. Variation in response to reward cues is captured in rats undergoing autoshaping experiments where the appearance of a lever precedes food delivery. Although no response is required for food to be delivered, some rats (goal-trackers) learn to approach and avidly engage the magazine until food delivery, whereas other rats (sign-trackers) come to approach and engage avidly the lever. The impulsive and often maladaptive characteristics of the latter response are reminiscent of addictive behaviour in humans, In a previous article, we developed a computational model accounting for a set of experimental data regarding sign-trackers and goal-trackers. Here we show new simulations of the model to draw experimental predictions that could help further validate or refute the model. In particular, we apply the model to new experimental protocols such as injecting flupentixol locally into the core of the nucleus accumbens rather than systemically, and lesioning of the core of the nucleus accumbens before or after conditioning. In addition, we discuss the possibility of removing the food magazine during the inter-trial interval. The predictions from this revised model will help us better understand the role of different brain regions in the behaviours expressed by sign-trackers and goal-trackers.

Keywords: Reinforcement Learning, Dopamine, Pavlovian conditioning, Autoshaping, Model-based, Model-free, Factored representation, Sign-tracker, Goal-tracker, Conditioned approach

1. Highlights

- We model goal-tracking and sign-tracking with a model-based/model-free combination
- We suggest that magazine in ITI is necessary for these distinct behaviours to emerge
- Phasic Dopaminergic activity can be explained by previously engaged features

2. Introduction

A significant number of models have been developed since the 1970s to describe Pavlovian and instrumental phenomena. Early models were mostly focusing on repro-

*Corresponding author
Email address: lesaint@isir.upmc.fr (Florian Lesaint)

ducing the averaged behaviour expressed within a population, neglecting inter-individual variations and possibly smoothing the true behaviour of individuals (Gallistel et al., 2004), or even masking the variation in behaviour. However, this variation is of particular interest when trying to identify those individuals within population prone to impulsive behaviours or having a higher risk of addiction (Flagel et al., 2011b; Saunders and Robinson, 2013; Huys et al., in press).

Recent studies have investigated such intervariability among rats undergoing an autoshaping experiment (Flagel et al., 2007, 2009, 2011b,a; DiFeliceantonio and Berridge, 2012; Mahler and Berridge, 2009; Robinson and Flagel, 2009; Meyer et al., 2012; Fitzpatrick et al., 2013), where a lever (conditioned stimulus, CS) was presented for 8 seconds, followed immediately by delivery of a food pellet (unconditioned stimulus, US) into an adjacent food magazine. Although no response was required to receive the reward, with training, some rats (sign-trackers; STs) learned to rapidly approach and engage the lever-CS. However, others (goal-trackers; GTs) learned to approach the food magazine upon CS presentation, and made anticipatory head entries into it. Some rats (intermediate group; IG) presented a mixed behaviour, switching between lever and magazine during presentation of the CS, and sometimes engaging both during one trial. Furthermore, in STs, phasic dopamine release in the core of the nucleus accumbens, measured with Fast Scan Cyclic Voltammetry (FSCV), matched the pattern that would be predicted by reward prediction error (RPE) signalling, and dopamine was necessary for the acquisition of a sign-tracking conditioned response (CR). In contrast, despite the fact that GTs acquired a Pavlovian conditioned approach response, this was not accompanied with the expected RPE-like dopamine signal, nor was the acquisition of a goal-tracking CR blocked by administration of a dopamine antagonist (see also Danna and Elmer (2010)). While the proportion of STs and GTs in the population varies (Fitzpatrick et al., 2013), both phenotypes are typically represented in an outbred population.

To our knowledge, only one model (Lesaint et al., 2014) accounts for these experimental results and has been validated with existing data. This model is built on a combination of model-free and model-based systems (Daw et al., 2005; Clark et al., 2012; Huys et al., in press) and extended with state factored representations. Combining multiple systems enables the model to express a large repertoire of behaviours and considering features within states enables the model to learn Pavlovian impetuses (Dayan et al., 2006) specific to the Pavlovian features within the task.

In this paper, we review the model described by Lesaint et al. (2014), extending it with a new tool to improve its reliability. We suggest new experimental protocols and some new analyses of the data that would further validate the model and strengthen its explanatory power, refine our understanding of the role of the nucleus accumbens in the described behaviours, and help clarify the impact of some choices made in the original protocol.

3. Material and methods

The model from which the present results are extracted is described in depth in a previous article (Lesaint et al., 2014). It is composed of two distinct reinforcement learning systems that collaborate to define the action to be selected at each step of the experiment (see Figure 1 A; Clark et al. (2012)).

The first system, a model-based system (MB), incrementally learns a model of the world (a transition function $\mathcal{T}$ and a reward function $\mathcal{R}$) from which it infers values ($\mathcal{A}$) for each action in each situation, given the classical following formulas:

$$\mathcal{Q}(s,a) \leftarrow \mathcal{R}(s,a) + \gamma \sum_{s'} \mathcal{T}(s'|s,a) \max_{a'} \mathcal{Q}(s',a') \quad (1)$$

$$\mathcal{A}(s,a) \leftarrow \mathcal{Q}(s,a) - \max_{a'} \mathcal{Q}(s,a') \quad (2)$$

where the discount rate $0 \leq \gamma \leq 1$ classically represents the preference for immediate versus distant rewards. At each step, the most valued action is the most rewarding on the long run (e.g. approaching the magazine to be ready to consume the food as soon as its delivery). It favours goal-tracking because this is the shortest path towards the rewarding state (see Figure 1 B).

The second system, a revised model-free system, learns values ($\mathcal{V}$) over features (e.g. food, lever or magazine). Contrary to the first system, which uses a classical abstract state representation, it relies on the features that compose these abstract states. In traditional reinforcement learning, each situation that can be encountered by the agent is defined as an abstract state (e.g. arbitrarily defined as s_1, s_2 ... s_x), such that similarities between situations (e.g. presence of a magazine) are lost. By using features, we reintroduce the capacity to use and benefit from these similarities. The second system is further defined as the feature model-free system (FMF). It relies on a RPE signal δ, computed as follows:

$$\begin{aligned} \mathcal{V}(c(s,a)) &\leftarrow \mathcal{V}(c(s,a)) + \alpha\delta \\ \delta &\leftarrow r + \gamma \max_{a'} \mathcal{V}(c(s',a')) - \mathcal{V}(c(s,a)) \end{aligned} \quad (3)$$

where $c : \mathcal{S} \times \mathcal{A} \rightarrow \{lever, magazine, food, \emptyset\}$ is a feature-function that returns the feature $c(s,a)$ the action a was focusing on in state s (e.g. it returns the lever when the action was to engage with the lever). We hypothesized that, similarly to classical model-free systems, δ parallels phasic dopaminergic activity (Schultz, 1998). This signal enables to revise and attribute values, seen as motivational, to features without the need of the internal model of the world used by the MB system. When an event is fully expected, there should be no RPE as its value is fully anticipated. When an event is positively surprising, there should be a positive RPE. Actions are then valued by the motivational value of the feature they are focusing on (e.g. engaging with the lever would be valued given the general motivational value of the lever). Hence, it favours actions that engage with the most motivational features. This might lead to favour suboptimal actions with regard to maximizing rewards (e.g. engaging with the lever keeps the rat away from the soon to be rewarded magazine). It favours sign-tracking (a suboptimal path, see Figure 1 B) as the lever, being a full predictor of reward, earns a strong motivational value relative to the magazine.

The model does not base its decision on a single system at a time, rather the values of the MB system ($\mathcal{A}_{MB}$) and the FMF system ($\mathcal{V}_{FMF}$) are integrated such that a single decision is made at each time step: producing a sort of cooperation between the two systems. The values computed by these two systems are then integrated through a weighted sum and passed to a softmax action selection mechanism that converts them into probabilities of selecting the action given a situation (see Figure 1 A). The integration is done as follows:

$$\mathcal{P}(s,a) = (1-\omega)\mathcal{A}_{MB}(s,a) + \omega\mathcal{V}_{FMF}(c(s,a)) \quad (4)$$

where $0 \leq \omega \leq 1$ is a combination parameter which defines the importance of each system in the overall model. Varying ω (while leaving the other parameters of the model unchanged) is sufficient to reproduce the characteristics of the different subgroups of rats (Lesaint et al., 2014). The previous experimental data could be reproduced by having STs give a stronger weight to the FMF system whereas having GTs give a stronger weight to the MB system. FMF and MB systems are then updated according to the action a taken by the full model in state s - even if the systems would have individually favoured different actions -and the resulting new state s' and retrieved reward r, as previously done in other computational models involving a cooperation between model-free and model-based systems (Caluwaerts et al., 2012).

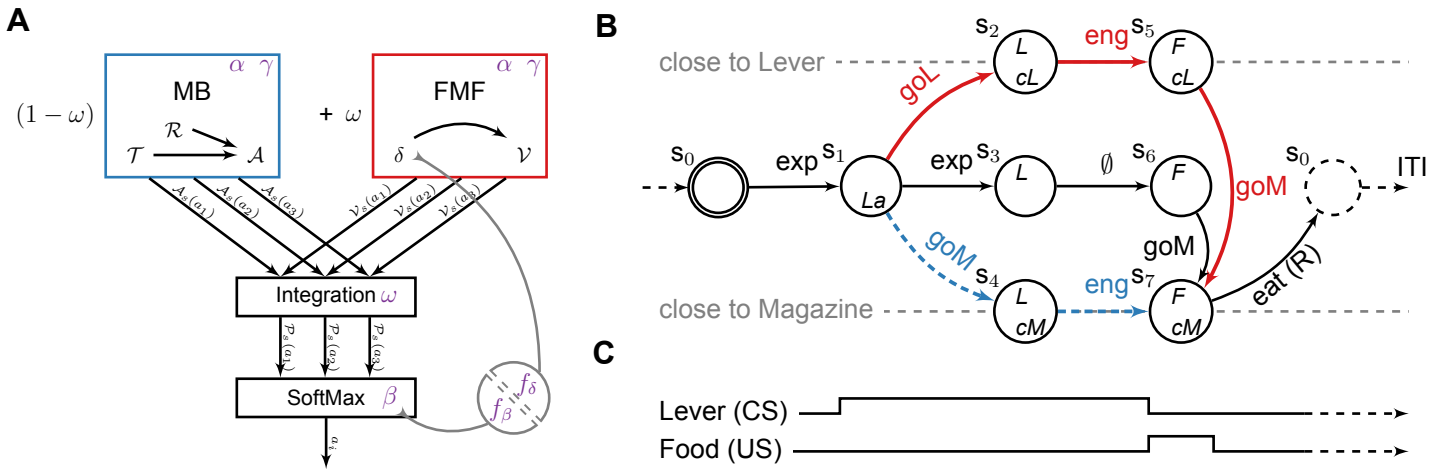

Figure 1: **Model and Markov Decision Process used for simulations. (A)** The model is composed of a model-based system (MB, in blue) and a Feature-Model-Free system (FMF, in red) which provide respectively an advantage function $\mathcal{A}$ and a value function $\mathcal{V}$ values for actions a_i given a state s. These values are integrated in $\mathcal{P}$, prior to be used into an action selection mechanism. The various elements may rely on parameters (in purple). The impact of flupentixol on dopamine is represented by a parameter f that influences the action selection mechanism and/or any reward prediction error that might be computed in the model. **(B)** MDP accounting for the experiments described in Flagel et al. (2009, 2011b); Robinson and Flagel (2009); Meyer et al. (2012). States are described by a set of variables: L/F - Lever/Food is available, cM/cL - close to the Magazine/Lever, La - Lever appearance. The initial state is double circled, the dashed state is terminal and ends the current episode. Actions are *eng*age with the proximal stimuli, *exp*lore, or *go* to the *M*agazine/*L*ever and *eat*. The path that STs should favour is in red. The path that GTs should favour is in dashed blue. **(C)** Time line corresponding to the unfolding of the MDP.

3.1. Simulations of experimental protocols

The experiment is described through an episodic Markov Decision Process (MDP) that represents one trial of the session (see Figure 1 B,C). The inter-trial interval (ITI), not being part of the MDP, is simulated between each run by revising downward the magazine value ($\mathcal{V}_{FMF}(M) \leftarrow (1 - u_{ITI}) \times \mathcal{V}_{FMF}(M)$, u_{ITI} being a parameter of the model). This simulates the hypothesis that the presence of the magazine in the absence of food delivery reduces its value. If the magazine were removed during ITI, we would expect no revision of its value.

The model is used to simulate experiments that involved injections of flupentixol, an antagonist of dopamine, either systemically or within the core of the nucleus accumbens. In the case of local injections, assuming that the FMF system relies on the core of the nucleus accumbens, we simulate the impact of flupentixol on phasic dopamine by degrading the reward predictions errors as follows:

$$\delta \leftarrow \begin{cases} \delta - f & \text{if } \frac{\delta - f}{\delta} \geq 0 \\ 0 & \text{otherwise} \end{cases} \tag{5}$$

where $0 \leq f < 1$ represents the impact of flupentixol. Its effect is defined such that flupentixol injections cannot lead to negative learning when RPE is positive, but at most blocks it. In the case of systemic injections, we also assume an additional impact on tonic dopamine (Humphries et al., 2012), which affects the action selection process. We simulate this impact by revising the temperature parameter ($\beta \leftarrow \frac{\beta}{1-f}$). Hence, flupentixol favours random exploration instead of using learned values to take a decision.

Some predictions presented here suggest to lesion the core of the nucleus accumbens. Such a lesion is simulated by removing the FMF system from the model , i.e. all values that would have come from the system are replaced by 0. The rest of the model is left intact. Equation 4 can

be replaced by:

$$\mathcal{P}(s,a) = (1-\omega)\mathcal{A}_{MB}(s,a) \quad (6)$$

3.2. Index Score

Introduced by Meyer et al. (2012), the Pavlovian Conditioned Approach (PCA) Index Score provides a metric to categorize rats as STs, GTs or IGs independent of the rest of the population. That is, instead of ordering rats based on their engagement with the lever and splitting the population in 3 groups of approximately equal size, as done in previous studies (Flagel et al., 2007; Robinson and Flagel, 2009), classifying rats based on PCA Index minimizes the chances of misclassification and allows one to compare across studies or populations of rats. The PCA Index relies on the number of contacts with the lever and the magazine, the probability to engage with one versus the other and the latencies to act towards each (Table 1 in (Meyer et al., 2012)).

We developed a similar Index Score as it provides a good metric for some of the predictions described here. Simulated rats whose score is > 0.5 are defined as STs. Simulated rats that have a score < -0.5 are defined as GTs. Remaining rats are defined as IGs. Table 1 explains how it is computed based on the last two sessions of simulations. Contrary to the PCA Index Score, it cannot use latencies as they are not accounted for by the model.

3.3. Estimation of model parameters

The model relies on a set of 8 parameters (a shared learning rate, a shared discount rate, a selection temperature, an integration parameter, a ITI impact parameter and 3 initial conditions) that need to be tuned for simulations to fit experimental data. We use the multi-objective algorithm NSGA-II (Deb et al., 2002; Mouret and Doncieux, 2010) to find the best values (solutions) for the parameters. This method is an efficient tool to fully explore the high dimensional parameter space and avoid local minima.

As in (Lesaint et al., 2014), we search a set of parameter values per group. The two first objectives of the fitness function are to fit the averaged behaviours of the simulated group to the averaged behaviours of the experimental group. More formally, for each group, we try to minimize the least square error between the probabilities of rats and simulated rats to engage with the magazine and the lever over time (see Table 2). This results in multiple solutions that are compromises between these two objectives. We subsequently select one of the solutions that is visually acceptable (no misclassification, and a good compromise between the two other criteria).

We noticed however, that without further constraints, as we are fitting averaged data, some of the resulting solutions could induce great variability of behaviour within a group, leading to misclassification. For example, a simulated rat classified as a GT by its parameters could have behaved as a ST and went undetected as its behaviour would have been diluted in the averaged behaviours of the simulated GT group.

The fitness function was extended with a new criterion based on the Index Score (see Table 2), to favour sets of parameter values that lead to groups of rats that did not introduce such errors, hence without strong inter variability. This is consistent with experimental data (Meyer et al., 2012). The resulting new sets of parameter values (see Table 3) did not affect the explanatory power of the model.

This metric ensures, for example, that using a set of parameter values for sign-tracking will produce a sign-tracker when applying the model in a simulation reproducing the original experiment. Interestingly, it allows us to predict qualitatively what the behaviour of such a rat (ST in normal conditions) would be in new experimental conditions: for example, whether the acquisition or the expression of the behaviour would be blocked or shifted to intermediate or even a goal-tracking behaviour, according to the Index Score defined above.

Response Bias(n)	$= (LeverPresses–MagazineEntries)/(LeverPresses + MagazineEntries)$
Probability Difference(n)	$= p(LeverPress)–p(MagazineEntry)$
Score(n)	$= [ResponseBias(n) + ProbabilityDifference(n)]/2$
Index Score	$= [Score(6) + Score(7)]/2$

Table 1: **Formulas for deriving the Index Score.** The Index Score provides a way to classify rats as STs, GTs or IGs, independently of the rest of the population. It relies on averaging scores computed for the last two sessions of the simulations. The Score for session n is derived by averaging its Response Bias and its Probability Difference. Responses Bias is a ratio between the difference in lever presses versus magazine entries and the total number of entries. Probability Difference is the difference between the probability to engage with the lever and the probability to engage with the magazine.

Objective	Formula
Best fit magazine engagement	$\min(\sum_{s_i \in \text{sessions}} (p^{s_i}_{Sim}(engM\|Group) - p^{s_i}_{Obs}(engM\|Group))^2)$
Best fit lever engagement	$\min(\sum_{s_i \in \text{sessions}} (p^{s_i}_{Sim}(engL\|Group) - p^{s_i}_{Obs}(engL\|Group))^2)$
Penalize parameters that lead to misclassification	$\min(\sum_{a_i \in \text{animals}} \|refPCA(Group) - IndexScore^{a_i}(Group)\|)$

Table 2: **Revised fitness function.** Lists of the multiple objective/criterion of the fitness function applied to each simulated group. *refPCA* is 1, 0 and -1 for STs, IGs, GTs respectively. This function is combined with NSGA-II to retrieve parameter values that best reproduce the experimental results. It results in a Pareto front of parameters from which we select by hand the solution that is consistent (no agent being misclassified) and that best visually fits the observed data (between engaging with the lever versus engaging with the magazine).

Type	ω	β	α	γ	u_{ITI}	$\mathcal{Q}_i(s_1, L)$	$\mathcal{Q}_i(s_1, \emptyset)$	$\mathcal{Q}_i(s_1, M)$
ST	0.501	0.243	0.027	0.946	0.845	0.263	0.272	0.344
IG	0.095	0.241	0.885	0.989	0.840	0.059	0.142	0.732
GT	0.081	0.063	0.033	0.483	0.893	0.936	0.022	0.099

Table 3: **Parameters used to produce the presented results.** All results were generated based on the same parameters. Some parameters might not be used or erased depending on the specific experimental protocol simulated.

4. Results

The model has already been validated on a set of behavioural, physiological and pharmacological data (Lesaint et al., 2014). Interestingly, while the model was only tuned to fit the behavioural data for each group, simulations of additional experiments without changing the parameters were consistent with the remaining experimental data.

The model accounts for the respective engagements of STs and GTs towards distinct specific features (Flagel et al., 2007, 2009, 2011b). It reproduces the difference in patterns of dopaminergic activity for GTs and STs (Flagel et al., 2011b). It also reproduces behaviours indicative of incentive salience attribution, including the conditioned reinforcement effect of the lever shown to a greater extent in STs than GTs (Robinson and Flagel, 2009), and the consumption-like engagement of the lever or magazine (Mahler and Berridge, 2009; DiFeliceantonio and Berridge, 2012). Finally, it also reproduces the impact of flupentixol injected either systemically prior to training (Flagel et al., 2011b), i.e. during acquisition, or locally after the rats have acquired their respective conditioned responses (Saunders and Robinson, 2012), i.e. expression.

In the following sections, taking inspiration from the set of studies used to validate the model, we generate predictions that new experiments or extended analyses of the data could confirm.

4.1. Dopaminergic patterns of activity

The model parallels the dopaminergic activity recorded in the core of the nucleus accumbens by Fast Scan Cyclic Voltammetry with the RPE signal used in the FMF system. At US time, the RPE signal within the FMF system

comes from the difference between the value of the previously engaged cue and the value of the delivered food. At CS time, it mainly reflects the value of the most rewarding cue between the lever and the magazine.

STs and GTs dopaminergic patterns at CS and US time are very distinct (Flagel et al., 2011b). While we observe a clear propagation of the signal from US to CS in STs (as expected from the classical RPE theory (Schultz, 1998)), this is not the case for GTs for which the CS and US signals are similar to one another and remain relatively constant across sessions (hence, in discrepancy with the classical theory).

In the model, the RPE signal is dependent of the feature previously focused on by the simulated rat. Thus, RPE patterns, averaged over sessions, strongly depend on the dominant path taken by the simulated rats before food delivery. Simulated STs, that mainly engage with the lever before food delivery, have an averaged signal that propagates from US to CS. This reflects that any rat that engages with the lever, eventually learns that it is a full predictor of food delivery. Simulated GTs, that mainly engage with the magazine before food delivery, have an averaged signal that do not show such a propagation. Indeed, the magazine is not fully informative of food delivery for any rat, hence a persistent reward prediction error remains at food delivery when engaging with the magazine during CS.

In Flagel et al. (2011b), recordings of dopaminergic activity in outbred rats were made to parallel those of the selectively bred STs and GTs but no recordings were made in outbred IGs. We would expect that IGs, whose behaviour fluctuate between sign-tracking and goal-tracking, would have a kind of mixed signal, averaging between those following from sign-tracking and goal-tracking. The current parameters values used in the model suggest that we would expect a high signal at CS time that would converge to a certain point, while at the meantime, the signal at US time would keep fluctuating without fully disappearing (see Figure 2).

Note that the visual results of this prediction are not identical with those in Lesaint et al. (2014). Contrary to ST and GT behaviours that deeply rely on the mechanisms, IG results strongly depend on the parameter values, which are significantly different with the introduction of the new score. Experimental recordings could help us refine the appropriate set of values for further predictions.

The initial analysis (Flagel et al., 2011b) and its reproduction (Lesaint et al., 2014) was done without taking into account the features engaged by animals prior to food delivery, possibly averaging very distinct patterns. The model predicts that if we were to organize the data per groups and actions rather than only per groups, we would observe patterns as shown in Figure 3. At the time the CS is presented, there should be no differences as all rats are exploring the world and not expecting the lever appearance, hence the positive RPE common to all rats. The difference would be at US time.

STs previously engaged with the lever (Figure 3 A) would show a classical propagation pattern, similar to the one of the initial analysis, as this condition dominates in the data. It reflects the fully predictive value of the lever. STs previously engaged with the magazine (Figure 3 C) would show a significant peak of DA activity, as they almost never engage with the magazine and hence attribute a low value to it, leading to an expected significant RPE.

GTs previously engaged with the magazine (Figure 3 D) would show an absence of propagation and patterns of DA activity that follow those at CS time, similar to the one of the initial analysis, as this condition dominates in the data. It reflects the difference between the value of the food delivered and the lower motivational value of the magazine. GTs previously engaged with the lever (Figure 3 B) would show a noisy dopaminergic activity that would decrease with time as the predictive value of the lever is learn.

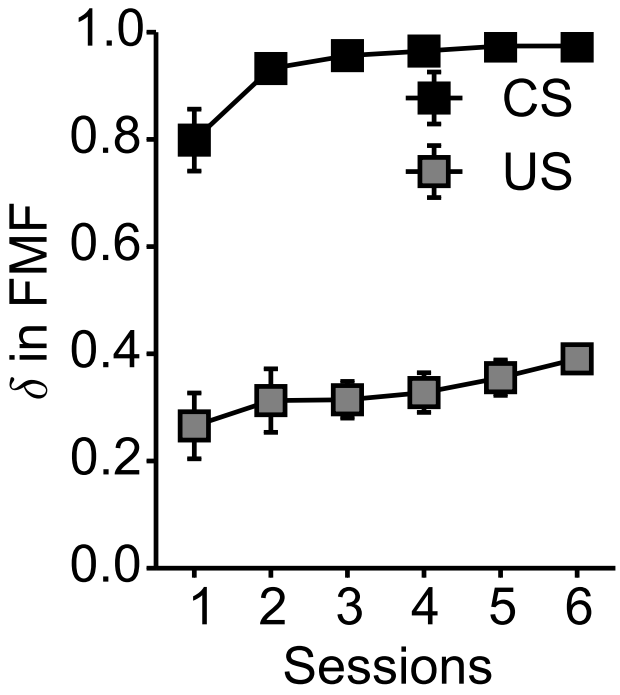

Figure 2: **Prediction of the model about expected patterns of dopaminergic activity in intermediate rats.** Data are expressed as mean ± SEM. Average RPE computed by the FMF system in response to CS and US presentation for each session of conditioning in the intermediate group.

4.2. Removal of magazine during the ITI

In the present model, the simulation of the ITI has a significant impact on the data. We hypothesize that the permanent presence of the magazine during the whole experiment lead animals to revise its associated motivational value, upward at lever retraction (i.e. food delivery) and downward during the ITI as there is no reward to be found then. Hence, on average, its presence does not guarantee access to food. In contrast, the time-locked presence of the lever before food delivery would lead to learn and maintain the motivational value of the lever to a certain level, as its presence guarantees food to be delivered.

First, by keeping the motivational value of the lever higher than that of the magazine in the FMF system, it makes simulated rats favouring this system (STs) to follow a sign-tracking policy. The small contribution of the MB system, which would attract rats towards the magazine does not compensate. Thus, the presence of the magazine in ITI is central for the emergence of STs in the model.

Second, by revising downward the magazine value between episodes, it maintains a discrepancy between the expectation (value) and the observation (reward) at food delivery in simulated rats being engaged with the magazine. This leads to the persistent positive RPE at US time and prevents a full propagation of the signal to CS time. Thus, the presence of the magazine in ITI is also central for the model, to explain the distinct dopaminergic patterns of activity in STs versus GTs that have been observed in Flagel et al. (2011b).

Third, we also hypothesize that values of the FMF system account for the motivational engagement, i.e. incentive salience, observed in rats towards either the lever or the magazine. The higher motivational value of the lever relative to that of the magazine implies that simulated rats

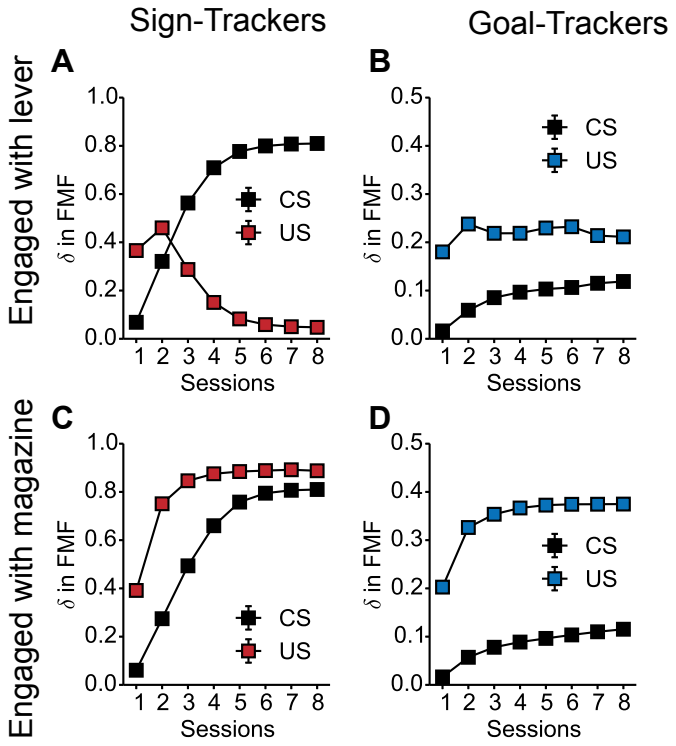

Figure 3: **Predictions about patterns of dopaminergic activity per groups and per actions.** Average RPE computed by the FMF system in response to CS and US presentation for each session of conditioning for STs **(A,C)** and GTs **(D,B)** when engaged with the lever **(A,B)** or with the magazine **(C,D)**. The model predicts very distinct patterns of activity depending on the feature engaged with prior to food delivery.

chew/bite more the lever than the magazine. While not central to the model, it is consistent with experimental observations (Mahler and Berridge, 2009; DiFeliceantonio and Berridge, 2012).

If no magazine were available during the ITI then, according to the model, the magazine would not loose its motivational value, as it would become a full predictor of food delivery and be highly valued. Hence, we would expect (1) an increased motivational engagement (chew/bite) towards the magazine, (2) a decreased tendency in sign-tracking within the population and (3) a different pattern of dopamine activity when goal-tracking for all rats.

As the motivational value of a feature accounts for the level of motivational engagement towards it, a higher motivational value of the magazine, relative to a control group, would necessarily lead to a relatively stronger motivational engagement towards it.

As the motivational value of the magazine would be as high as that of the lever, there should be no reason for rats relying mainly on the FMF system (STs) to favour one over the other, hence shifting to behaviours similar to those of IGs and GTs (see Figure 4). GTs, relying mainly on the MB system would not be deeply affected (see Figure 4 B).

Finally, as the presence of the magazine would be time-locked to the moments before the delivery of food, we would expect a propagation of the dopamine signal from US time to CS time (see Figure 5). At some point (after the value of the food has been fully learned) the signal at US time should start decreasing. Note that if we would have used the same parameters (except for the weighting parameter) to simulate STs and GTs, we would have ex-

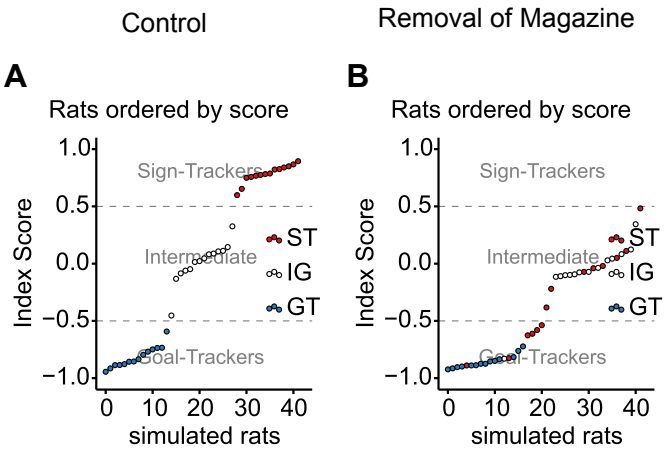

Figure 4: **Distribution of rats given the removal of the magazine during the ITI.** Simulated rats ordered by their Index Score. In blue simulated rats using parameter values tuned for GTs, in red for STs and in white for IGs in the classical condition (control). Rats with a score < -0.5 are GTs, with a score > 0.5 are STs and remaining rats are IGs. **(A)** As expected, rats using parameters' value for GTs are classified as GTs. Same for STs and IGs. **(B)** Without magazine during the ITI, simulated rats that would have been classified as GTs in normal conditions are still classified GTs. However, rats that would have been classified as STs (red) have a score that classify them as GTs or IGs. One IG (white) is now classified as GT.

pected an identical RPE signal for STs and GTs, and we know this is not the case based on existing data (Flagel et al., 2011b).

The expected decreased tendency in sign-tracking within the simulated population does not mean that simulated rats would not be attracted any more by the lever. Simulated rats would indeed be attracted by both the lever and magazine because their FMF system attributes a high motivational value to all signs preceding reward delivery. Combined with the contribution of the MB system which attracts rats towards to magazine, it could make the simulated animal engage more with the magazine than with the lever. Thus if the computational model is valid, this would mean that the tendency to sign-track in real animals can be gradually changed by affecting some of the signs or features present in the context of the task (here the magazine during the ITI).

4.3. Injections of flupentixol in the core of the nucleus accumbens

In the model, flupentixol, an antagonist of dopamine, is hypothesized to impact the RPE (hypothesized to parallel phasic dopamine) used in the FMF system, putatively based within the core of the nucleus accumbens. Flupentixol is also assumed to affect any action selection process, relying on tonic dopamine (Humphries et al., 2012). Hence, under systemic injections of flupentixol, the learning process of the FMF system is disrupted and actions are almost randomly picked barely using learned values.

With systemic injections of flupentixol (Flagel et al., 2011b), no goal-tracking nor sign-tracking is expressed in the population. However, when afterwards released from flupentixol, GTs fully express goal-tracking, whereas STs behave as untrained rats.

The model accounts for the absence of behaviours under flupentixol by the hypothesized impact of flupentixol on the action selection process, blocking the expression of any acquired behaviour Lesaint et al. (2014). The subsequent absence of sign-tracking on a last session free of flupentixol is explained by the disruption of the FMF system during the 7 first sessions, blocking behaviour acquisition. The full expression of goal-tracking as soon as flupentixol is removed, relies on the unaffected learning process in

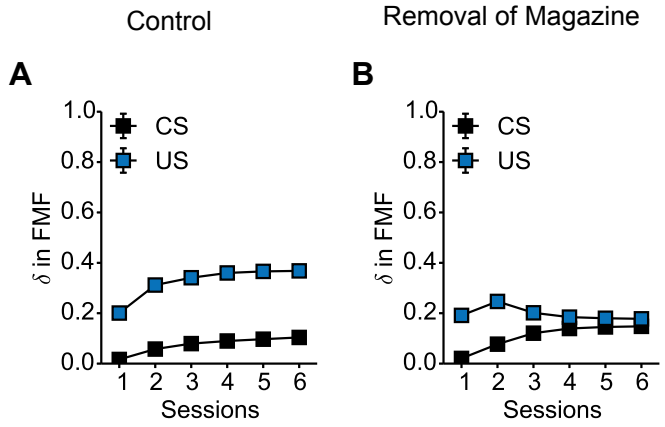

Figure 5: **Patterns of dopaminergic activity for GTs given the removal of the magazine during ITI.** Average RPE computed by the FMF system in response to CS (black) and US (blue) presentation for GTs for each session of conditioning. It is hypothesized to parallel the patterns of dopaminergic activity observed by FSCV in the core of the nucleus accumbens. **(A)** With the classical protocol (control), signal at CS and US seems to follow similar trends and there is no propagation of signal from US to CS. **(B)** When magazine is time-locked to CS presentation, the value of US is propagated to the CS. Thus, the signal at US time, after sufficient learning (2 first sessions) start decreasing in favour of the CS time.

the MB system, assumed to be dopamine-independent and hence keeps learning under flupentixol, but which values are simply not used by the softmax function.

The model predicts that if flupentixol were injected locally in the core of the nucleus accumbens rather than systemically prior to acquisition, GTs would normally express their behaviour, as the action selection mechanism would not be disrupted and make use of the values learned in the MB system; whereas STs' behaviour would remain blocked because of the disruption of the FMF system (see Figure 6), and this is indeed what happend when Saunders and Robinson (2012) locally injected flupetixol after the behaviours were already acquired.

4.4. Lesions of core of the nucleus accumbens

While we did not try to find all anatomical counter parts of the mechanisms involved in the model, the hypothesis that the FMF system relies mainly on the core of the nucleus accumbens is important for the model. Indeed, RPEs used in the FMF system are compared with the dopaminergic recordings (using FSCV) in the core of the nucleus accumbens. As already stated, the values learned by the FMF system are a key component in the emergence of sign-tracking behaviours within a population and assumed to reflect the motivational engagement observed towards the magazine and the lever.

As stated in the previous section, Flagel et al. (2011b) studied the impact of systemic injection of flupentixol on the acquisition of sign-tracking and goal-tracking. They observed that the acquisition of a goal-tracking behaviour did not require a fully functional dopaminergic system contrary to sign-tracking. Another study (Saunders and Robinson, 2012) focused on the impact of local injections of flupentixol in the core of the nucleus accumbens on the expression of sign-tracking and goal-tracking, after 8 days of conditioning. On the last day, with a sufficient dose of flupentixol, they observed a decrease in the general tendency to sign-track in the overall population while leaving the level of goal-tracking unaffected.

Simulating injections of flupentixol in the core of the nucleus accumbens, by disrupting RPEs in the FMF system and hence its contribution in the decision, the model accounts for these last observations. The action selection

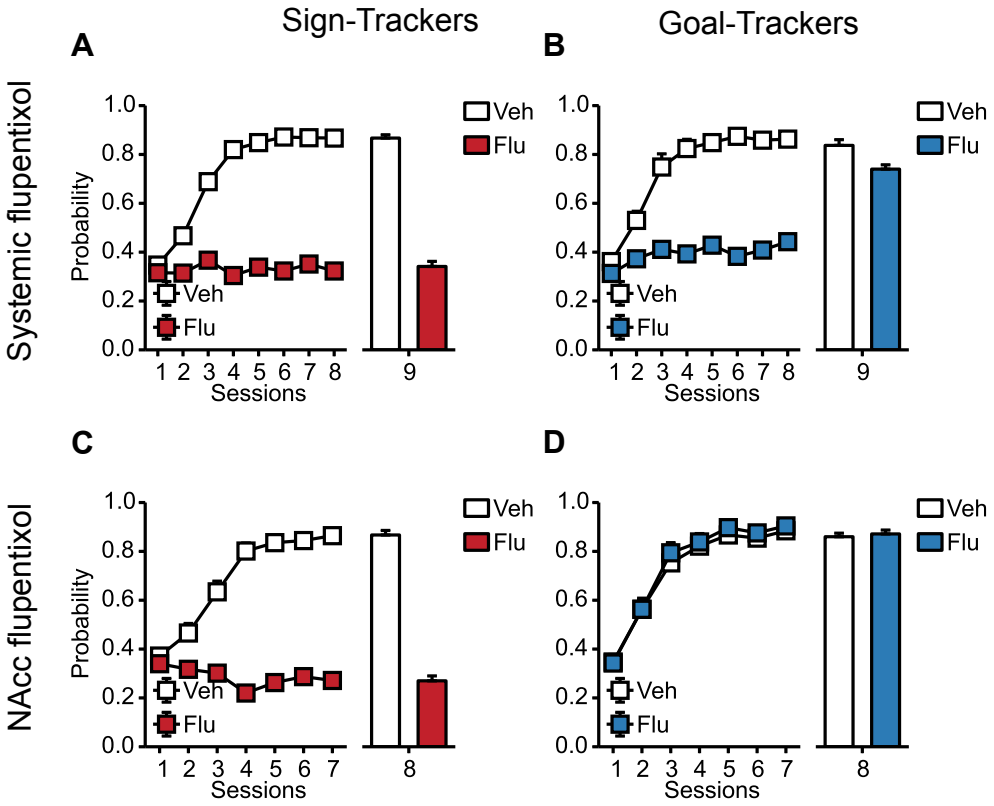

Figure 6: **Simulated impact of injections of flupentixol. (A,B)** Simulation of the impact of systemic injections of flupentixol (Flagel et al., 2011b) on the probability to approach the lever for STs **(A)** and the magazine for GTs **(B)** during lever presentation. Last session is without flupentixol. Under flupentixol (first 7 sessions), both sign-tracking and goal-tracking are blocked. On the last flupentixol-free session (8 session), STs are unable to express sign-tracking, its learning having been blocked, whereas GTs fully express goal-tracking, which learning was only covert. **(C,D)** Simulation of the impact of local injections of flupentixol in the core of the nucleus accumbens, hypothesised to impact only the FMF-system. Contrary to the initial experiment, the injections being localized to the FMF-system, the action selection mechanism is not impacted. Hence, GTs fully express goal-tracking during the first 7 sessions **(C)**. STs are still unable to express sign-tracking **(D)**.

mechanism remains functional and makes use of the MB system values, such that the behaviour of GTs is preserved while the one of STs is disturbed and leads to a decrease in sign-tracking in the overall population.

We expect that lesions of the core of the nucleus accumbens would lead to similar effects as the above experiments.

Lesions of the core of the nucleus accumbens *prior* to the experiment would (1) block the expression of sign-tracking responses and (2) stop the motivational engagement towards the magazine or the lever during approaches.

By disabling the FMF system (setting and keeping all values to 0), it cannot favour the lever over the magazine any more. STs would therefore act randomly , approaching lever and magazine indifferently, as observed in IGs. We would expect a shift towards goal-tracking similar to the one expected for removing the magazine during the ITI (as in Figure 4).

However, while a magazine removal would lead to an increase in motivational engagement, we expect such a lesion to block any consumption-like behaviour. Especially, we would expect GTs' approach behaviour to remain similar to control group, but without subsequent chewing and biting of the magazine.

We would expect that lesions of the core of the nucleus accumbens *after* the experiment would disrupt the ten-

dency to sign-track in the overall population, while leaving the tendency to goal-track intact (see Figure 7). However, contrary to flupentixol injections, that needed 35 min of infusion for a visible effect, we would expect the effect to be immediate with a lesion. Such a lesion would disrupt the FMF system, hence (1) suppressing any consumption-like engagement towards the features (motivational values being kept to 0), and (2) stop favouring engagements towards the lever. The lesion would leave the MB system unaffected and have no impact on the general tendency to goal-track.

5. Discussion

Relying on a model that was previously validated using experimental data to account for variability in rats undergoing an autoshaping paradigm (Lesaint et al., 2014), we generate an additional set of behavioural, physiological and pharmacological predictions.

We predict that dopaminergic patterns for IGs should be a mixed signal between those observed for STs and GTs. We predict that looking separately at the DA patterns given the prior engagement towards either the lever or the magazine should lead to clearly distinct patterns. We predict that the removal of the magazine during the ITI should lead to an increased motivational engagement towards the magazine, a decreased tendency in sign-tracking within the population and a different pattern of dopaminergic activity when goal-tracking. Finally, we predict that local injections of flupentixol to the core of the nucleus accumbens would preserve goal-tracking and prevent the learning of a sign-tracking response, a result that should also be observed following lesions of the core of the nucleus accumbens prior to conditioning. Lesions after conditioning, would only block the expression of the learned sign-tracking behaviour.

An important limitation of the present predictions is that most of them are based on the behaviour that is expected to emerge from naive rats trained in a revised protocol, assuming that they would have behaved in a specific manner in the standard protocol (e.g. expecting a supposed ST to goal-track). To overcome this difficulty, one must look at the population level rather than the individual level (Saunders and Robinson, 2012), which might be problematic as the proportion of GTs, STs and IGs is highly variable in a population (Meyer et al., 2012; Fitzpatrick et al., 2013). An alternative would be to use selectively bred rats that can more or less be ensured to behave as STs or GTs in experimental conditions (Flagel et al., 2011b).

Another limit of the present predictions are the hypotheses on which they are based. It cannot be excluded that the core of the nucleus accumbens also contributes to the MB system, but not by its dopaminergic activity (Khamassi and Humphries, 2012; van Der Meer and Redish, 2011; McDannald et al., 2011) (but see van der Meer et al. (2010); Bornstein and Daw (2012); Penner and Mizumori (2012)). Hence, completely disrupting it might unexpectedly affect goal-tracking. Validating these predictions would help to confirm this hypothesis. In the initial model (Lesaint et al., 2014) we interpreted the parameter which simulates the ITI as accounting for the engagement of the rats towards the magazine during the ITI. Preliminary analyses of experimental data (not shown), while still inconclusive, tend to mitigate such a strong hypothesis. Hence, in the current article, we only assume that the presence of the magazine during the ITI impacts its general motivational value within the experiment. Validating such predictions would definitely help to clarify the impact of the ITI context on the expressed behaviours.

One could argue that, to some extent, describing STs with a MF system and GTs with a MB system could be sufficient to explain dominant behaviours (Clark et al., 2012). However, it would fail to explain the full and continuous spectrum of observed behaviours (Meyer et al., 2012). If the predictions that we make about IGs (which have an intermediate behaviour between STs and GTs)

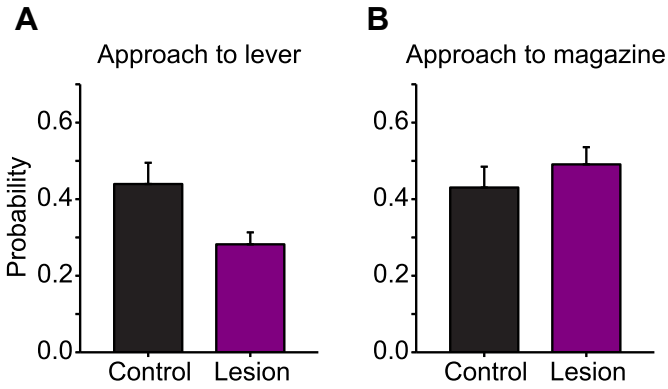

Figure 7: **Predictions of the impact of lesions of the core of the nucleus accumbens after conditioning.** General tendencies to sign-track **(A)** and goal-track **(B)** in a population of rats after training. Lesion of the core of the nucleus accumbens is simulated by a blockade of the FMF system in the model. We expect a decrease in the tendency to sign-track **(A)** with a lesion (purple) relative to a control group (black). General goal-tracking tendencies should remain unchanged **(B)**.

are correct, this would argue in favour of a continuum in the weighting between MB and MF systems rather than a pure dichotomy.

An alternative to the collaboration of both systems (through a weighted sum) would be a reciprocal inhibition, such that only one system would be working at a time. This would be sufficient to account for the previous point and may even be able to account for the absence of RPE pattern in the dopamine signal measures in GTs (Flagel et al., 2011b) without requiring a revision of the magazine value during ITI. The inhibition of the MF system in GTs would indeed prevent any RPE signal from being observed. However, it would be unable to properly account for the consumption-like engagement observed in both STs and GTs without some kind of extension (see Zhang et al. (2009) for a computational model of incentive salience). It would also fail to explain why the pharmacological disruptions of one system does not seem to let the other take control (Flagel et al., 2011b).

Another possibility would be that the two systems run in parallel but that only one is used to make the decision during a trial. Assuming that one system leads to the lever and the other to the magazine, we would expect IGs to behave as STs when engaging with the lever and GTs when engaging with the magazine. Experimental data goes against such interpretation. Meyer et al. (2012) observed that contrary to STs or GTs, IGs tend to approach both the magazine and the lever during single trials. Some rats even hold on to the lever while putting their head into the magazine (which no model that selects a single action at a time can reproduce). While the task representation does not allow multiple engagements in a trial, this suggests that both systems are active and contribute actively to their behaviour at all time. We would also expect rats to behave differently when using one system over the other, such that, for example, rats would actively engage with the lever but quietly wait in front of the magazine, which is not the case. Finally, the recent literature seems consistent with multiple systems working in parallel and partially contributing to a global decision (e.g. Daw et al. (2011)). Hence, this does not suggest take-over competition between the systems. Trial-by-trial analyses (Daw, 2011) would allow us to definitely rule out such alternatives. Finally, if only the output of the MF system was inhibited, given that the lever appearance is fully predictive of food delivery, no classical MF system

(relying on classical state representation) would reproduce the differences observed in phasic dopaminergic patterns between STs and GTs nor explain the differences of focused features. Hence, the model suggests to take features into consideration.

The interest of the current computational model lies in its combination of simple concepts actively used and accepted in the current field (Dual reinforcement learning and factored representations) but rarely used together, to account for a variability of experimental data, without resorting to arbitrary additions. As a result the current model does not behave as state of the art algorithms would on the same task and produces a suboptimal behaviour. This suboptimal behaviour is, however, in accordance with behavioural observations in rats.

Subsequent studies could benefit from a different approach to estimate parameters. We are currently fitting the model on the behavioural data per sessions and groups, using trial-by-trial analyses could prove a better tool to fit the parameters at the individual level (Daw et al., 2011) and comfort some choices in the architecture of the model.

It has been suggested that individuals for whom cues become powerful incentives (i.e. STs) are more prone to develop addiction (Saunders and Robinson, 2012). Thus, the current model and its predictions will allow us to further investigate and possibly identify the neural mechanisms that underlie addiction and related disorders. For example, the current model predicts that some manipulations could alter the behaviour of STs towards that of GTs, and the neurobiological targets of these manipulations may be used to alter drug-cue dependency and prevent relapse (For further discussion regarding the role of learning-related dopamine signals in addiction vulnerability, see Huys et al. (in press)).

To conclude, the current article refines the model previously described by Lesaint et al. (2014) with an additional metric that strengthens its explanatory power. It mainly suggests a set of predictions with which to further confront the model. The new proposed experiments would help to better localize the anatomical counterparts of the mechanisms involved and disentangle their contributions to the observed behaviours. It would also help in refining the hypotheses and simplifications of the model and we hope would confirm the interest and necessity of considering the features rather than the general situations encountered by rats when modelling this kind of phenomena.

Acknowledgments

References

Bornstein, A. M., Daw, N. D., 2012. Dissociating hippocampal and striatal contributions to sequential prediction learning. Eur J neurosci 35 (7), 1011–1023.

Caluwaerts, K., Staffa, M., N'Guyen, S., Grand, C., Dollé, L., Favre-Félix, A., Girard, B., Khamassi, M., 2012. A biologically inspired meta-control navigation system for the psikharpax rat robot. Bioinspiration & biomimetics 7 (2), 025009.

Clark, J. J., Hollon, N. G., Phillips, P. E. M., 2012. Pavlovian valuation systems in learning and decision making. Curr Opin Neurobiol 22 (6), 1054–1061.

Danna, C. L., Elmer, G. I., 2010. Disruption of conditioned reward association by typical and atypical antipsychotics. Pharmacol Biochem Behav 96 (1), 40–47.

Daw, N. D., 2011. Trial-by-trial data analysis using computational models. In: Delgado, M. R., Phelps, E. A., Robbins, T. W. (Eds.), Decision Making, Affect, and Learning: Attention and Performance XXIII. Vol. 23. Oxford University Press, Ch. 1.

Daw, N. D., Gershman, S. J., Seymour, B., Dayan, P., Dolan, R. J., 2011. Model-based influences on humans' choices and striatal prediction errors. Neuron 69 (6), 1204–1215.

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

Dayan, P., Niv, Y., Seymour, B., Daw, N. D., 2006. The misbehavior of value and the discipline of the will. Neural Netw 19 (8), 1153–1160.

Deb, K., Pratap, A., Agarwal, S., Meyarivan, T., 2002. A fast and elitist multiobjective genetic algorithm: Nsga-ii. IEEE Trans Evol Comput 6 (2), 182–197.

DiFeliceantonio, A. G., Berridge, K. C., 2012. Which cue to 'want'? Opioid stimulation of central amygdala makes goal-trackers show stronger goal-tracking, just as sign-trackers show stronger sign-tracking. Behav Brain Res 230 (2), 399–408.

Fitzpatrick, C. J., Gopalakrishnan, S., Cogan, E. S., Yager, L. M., Meyer, P. J., Lovic, V., Saunders, B. T., Parker, C. C., Gonzales, N. M., Aryee, E., 2013. Variation in the form of pavlovian conditioned approach behavior among outbred male sprague-dawley rats from different vendors and colonies: Sign-tracking vs. goal-tracking. PloS ONE 8 (10), e75042.

Flagel, S. B., Akil, H., Robinson, T. E., 2009. Individual differences in the attribution of incentive salience to reward-related cues: Implications for addiction. Neuropharmacology 56, 139–148.

Flagel, S. B., Cameron, C. M., Pickup, K. N., Watson, S. J., Akil, H., Robinson, T. E., 2011a. A food predictive cue must be attributed with incentive salience for it to induce c-fos mRNA expression in cortico-striatal-thalamic brain regions. Neuroscience 196, 80–96.

Flagel, S. B., Clark, J. J., Robinson, T. E., Mayo, L., Czuj, A., Willuhn, I., Akers, C. A., Clinton, S. M., Phillips, P. E. M., Akil, H., 2011b. A selective role for dopamine in stimulus-reward learning. Nature 469 (7328), 53–57.

Flagel, S. B., Watson, S. J., Robinson, T. E., Akil, H., 2007. Individual differences in the propensity to approach signals vs goals promote different adaptations in the dopamine system of rats. Psychopharmacology 191 (3), 599–607.

Gallistel, C. R., Fairhurst, S., Balsam, P., 2004. The learning curve: Implications of a quantitative analysis. Proceedings of the national academy of Sciences of the united States of america 101 (36), 13124–13131.

Humphries, M. D., Khamassi, M., Gurney, K., 2012. Dopaminergic control of the exploration-exploitation trade-off via the basal ganglia. Front Neurosci 6 (9).

Huys, Q. J. M., Tobler, P. N., Hasler, G., Flagel, S. B., in press. The role of learning-related dopamine signals in addiction vulnerability. Prog Neurobiol.

Khamassi, M., Humphries, M. D., 2012. Integrating cortico-limbic-basal ganglia architectures for learning model-based and model-free navigation strategies. Front Behav Neurosci 6 (79).

Lesaint, F., Sigaud, O., Flagel, S. B., Robinson, T. E., Khamassi, M., 2014. Modelling individual differences in the form of pavlovian conditioned approach responses: A dual learning systems approach with factored representations. PLoS Comput Biol 10 (2), e1003466.

Mahler, S. V., Berridge, K. C., 2009. Which cue to "want?" Central amygdala opioid activation enhances and focuses incentive salience on a prepotent reward cue. J Neurosci 29 (20), 6500–13.

McDannald, M. A., Lucantonio, F., Burke, K. A., Niv, Y., Schoenbaum, G., 2011. Ventral striatum and orbitofrontal cortex are both required for model-based, but not model-free, reinforcement learning. J Neurosci 31 (7), 2700–2705.

Meyer, P. J., Lovic, V., Saunders, B. T., Yager, L. M., Flagel, S. B., Morrow, J. D., Robinson, T. E., 2012. Quantifying individual variation in the propensity to attribute incentive salience to reward cues. PLoS ONE 7 (6), e38987.

Mouret, J.-B., Doncieux, S., 2010. SFERESv2: Evolvin' in the Multi-Core World. In: WCCI 2010 IEEE World Congress on Computational Intelligence, Congress on Evolutionary Computation (CEC). pp. 4079–4086.

Penner, M. R., Mizumori, S. J. Y., 2012. Neural systems analysis of decision making during goal-directed navigation. Progress in neurobiology 96 (1), 96–135.

Robinson, T. E., Flagel, S. B., 2009. Dissociating the predictive and incentive motivational properties of reward-related cues through the study of individual differences. Biol psychiatry 65 (10), 869–873.

Saunders, B. T., Robinson, T. E., 2012. The role of dopamine in the accumbens core in the expression of pavlovian-conditioned responses. Eur J neurosci 36 (4), 2521–2532.

Saunders, B. T., Robinson, T. E., 2013. Individual variation in resisting temptation: implications for addiction. Neurosci Biobehav Rev 37 (9), 1955–1975.

Schultz, W., 1998. Predictive reward signal of dopamine neurons. J Neurophysiol 80, 1–27.

van der Meer, M., Johnson, A., Schmitzer-Torbert, N. C., Redish, A. D., 2010. Triple dissociation of information processing in dorsal striatum, ventral striatum, and hippocampus on a learned spatial decision task. Neuron 67 (1), 25–32.

van Der Meer, M., Redish, A. D., 2011. Ventral striatum: a critical look at models of learning and evaluation. Curr Opin Neurobiol 21 (3), 387–392.

Zhang, J., Berridge, K. C., Tindell, A. J., Smith, K. S., Aldridge, J. W., 2009. A neural computational model of incentive salience. PLoS computational biology 5 (7), e1000437.