The Successor Representation as a model of behavioural flexibility

Alexis Ducarouge, Olivier Sigaud

Sorbonne Universités, UPMC Univ Paris 06, CNRS UMR 7222,
Institut des Systèmes Intelligents et de Robotique, F-75005 Paris, France
olivier.sigaud@isir.upmc.fr +33 (0) 1 44 27 88 53

Abstract :

Accounting for behavioural capabilities and flexibilities experimentally observed in animals is a major issue in computational neurosciences. In order to design a comprehensive algorithmic framework for this purpose, the model-free and model-based reinforcement learning (RL) components are generally taken as reference either in isolation or in combination. In this article, we consider the RL Successor Representation (SR) approach as an alternative. We compare it to the standard model-free and model-based models on three relevant experimental data-sets. These modelling experiments demonstrate that SR is able to account better for several behavioural flexibilities while being algorithmically simpler.

Mots-clés : Reinforcement learning, successor representation, behavioural flexibility, latent learning, internal motivation, policy revaluation.

1 Introduction

Reinforcement learning (RL) is a comprehensive framework that accounts for behavioural and neuronal data of animals addressing temporal difference learning tasks [11]. Because of historical perspectives in psychology [45, 43] and of the somewhat natural division of the RL field, much of the past RL modelling work was striving to disentangle if some aspects of animal behaviour fall under "model-free" or "model-based" mechanisms. The natural reaction to limitations of these quite antagonist systems was to combine them in a dual framework, requiring even an arbitration mechanism as third component [15, 23, 20, 31].

In order to advocate for this algorithmic perspective, the dual system theory has benefited from a lot of experimental works to find potential correlations between neuronal activations and model insights [35, 16, 7, 1, 15, 23, 49]. Nevertheless, even though plenty of correlations have been found (overall, in ventral striatum for reward prediction error of the model-free component and in lateral prefrontal cortex and infraparietal sulcus for state prediction error of the model-based component), results remain really puzzling in so far as there is a lot of crossed significant activations between the two systems which are generally assumed independent [6]. Furthermore, while dopaminergic activity is assumed to correlate with the reward prediction error update of model-free learning, more recent studies measuring dopamine levels show that this activity is much more present during phases in which the model-based system is supposed to be dominant [12]. This essentially challenges the idea that model-based system is inherently not based on temporal-difference learning, and thus that it does not rely on dopaminergic mechanisms. Furthermore, there exists numerous other theories about the role of dopamine which might also be reasonable and which may be inconsistent with the standard reward prediction error interpretation [4, 36].

Given these contradictory findings, some of the researchers supporting these dual system theories have started to advocate for a much more integrated perspective of these two systems [6, 20, 31, 13, 49]. Nonetheless, how they are supposed to interact, and the role of dopaminergic circuitry in such a dual system remain unclear [5].

In this context, the Successor Representation (SR) [8] recently emerged as a promising alternative [51, 10, 14, 5, 39, 2, 29, 34, 22]. The main feature of this approach is that it factors the temporal component of the task into a statistical contingencies RL problem, remodeling the initial estimation problem into a

straightforward reward evaluation which can take into account internal goals of the agent [8]. The learned representation of the task could be viewed as a partial internal model, or cognitive map of it environment [39] which is learned following a model-free fashion using a temporal difference learning rule. This map is then used in an inexpensive way to compute optimal policy given the reward structure of the task. Section 2 of the article introduces the method more formally.

The SR framework is endowed with complementary properties of classical model-based and model-free algorithms. Its learning mechanisms are as biologically plausible and computationally cheap as model-free one; and it allows several behavioural abilities and learning flexibilities which are often considered as sole privileges of model-based one [6, 34], while being simpler. Hence, we investigate in this article if the Successor Representation approach could explain the behavioural data of three classical experiments which emphasize some behavioural abilities as latent learning [3, 33, 45, 19], learning flexibility as immediate reward-policy revaluation [6] and the role of a changing motivation in learning [24, 17, 45]. We compare its capability to quantitatively fit the corresponding experimental data to two model-free (SARSA(λ)) and model-based algorithms as well as a hybrid theory in which both algorithms are computed in parallel and agents can use a weighted combination of the learned action-values to decide what to do [15, 6]. The SR algorithm distinguishes itself by its simplicity and is sometimes the only approach which can account for the observed behaviours.

2 Background and Methods

In this section we introduce the standard RL framework, reminding the model model-free and model-based approaches before presenting the Successor Representation algorithm and its eligibility trace extension.

2.1 RL background

A Markov Decision Process consists of a set of states $\mathcal{S}$, a set of actions $\mathcal{A}$, a discount factor γ, a reward function $\mathcal{R} : \mathcal{S} \rightarrow \mathbb{R}$, and a transition distribution $\mathcal{T} : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \rightarrow [0, 1]$ specifying the probability of transitioning to state $s' \in \mathcal{S}$ from state $s \in \mathcal{S}$ given action $a \in \mathcal{A}$. An agent chooses actions at each discrete time-step according to a stochastic policy $\pi : \mathcal{S} \times \mathcal{A} \rightarrow [0, 1]$.

The goal of the agent is to learn an optimal policy π^* that maximizes the expected cumulative discounted future *value* defined as $V(s_t) = \mathbb{E}_\pi[\sum_{k=t}^{\infty} \gamma^{k-t} R(s_k)]$.

2.2 Model-free and model-based approaches

The classical model-free and model-based approaches rely on different mechanisms; from an algorithmic standpoint the former gradually estimates and caches V directly from sample paths, without building any model of the structure of the MDP. Equivalently, the state-action expected cumulative discounted future *value* Q is generally preferred, defined as

$$Q^\pi(s_t, a_t) = \mathbb{E}_{a_{i>t} \sim \pi} \left[\sum_{k=t}^{\infty} \gamma^{k-t} R(s_k) | s_t, a_t \right]. \quad (1)$$

The classical off-policy temporal difference error used in Q-learning to iteratively update Q^π is:

$$\delta = R(s_t) + \gamma \max_a [Q^\pi(s_{t+1}, a)] - Q^\pi(s_t, a_t).$$

The model-based approach rather progressively estimates a model of the environment, namely the transitions $\mathcal{T}$ and rewards $\mathcal{R}$, from sample paths. It then recomputes at each time-step V or Q using either a form of tree search or dynamic programming to infer the optimal policy [41].

2.3 The Successor Representation

The Successor Representation (SR) approach is an approach to RL based on estimating state occupancy relations in a given environment according to the current policy [50, 40, 8, 9]. These relations are stored in a

SR matrix which encodes the expected cumulative discounted future state occupancy M^π of state $\bar{s}$ given a policy π followed from the initial state s_t given the initial action a_t:

$$M^\pi(s_t, \bar{s}, a_t) = \mathbb{E}_{a_{i>t}\sim\pi}\left[\sum_{k=t}^{\infty} \gamma^{k-t} \mathbb{1}[s_k = \bar{s}] | s_t, a_t\right],$$

where $\mathbb{1}[\cdot] = 1$ if its argument is true and 0 otherwise. It implicitly captures contingencies between states given the statistical structure of the MDP and the current policy.

A coefficient of the M^π matrix can be understood as encoding the discounted probability to reach a given state in the future, knowing that the agent takes a given action from a given state and then follows policy π.

As in other RL approaches, several Bellman recursive equations can be defined, one for each future state $\bar{s}$ occupancy prediction problem:

$$M^\pi(s_t, \bar{s}, a_t) = \mathbb{1}[s_t = \bar{s}] + \gamma \mathbb{E}_{a_{i>t}\sim\pi}[M^\pi(s_{t+1}, \bar{s}, a_{t+1})]. \tag{2}$$

One can use this equation to derive the state occupancy temporal difference (TD) error which is the grounding component of the incremental learning algorithm of the SR. Hence, we use the state occupancy TD error to specify the $|\mathcal{S}|$ on-policy update rules of an estimate of the SR - which we simply denote as M:

$$\forall \bar{s} \in \mathcal{S}, \quad M(s_t, \bar{s}, a_t) \leftarrow M(s_t, \bar{s}, a_t) + \alpha(\mathbb{1}[s_t = \bar{s}] + \gamma M(s_{t+1}, \bar{s}, a_{t+1}) - M(s_t, \bar{s}, a_t)).$$

Once the SR is learned, it can be used to infer the action value function defined in (1). Indeed, the expected cumulative discounted future value of choosing action a_t in state s_t given the policy π is given as

$$Q^\pi(s_t, a_t) = \sum_{\bar{s}\in S} M(s_t, \bar{s}, a_t)\widetilde{R}(\bar{s}) \tag{3}$$

where $\widetilde{R}$ is the estimated immediate reward in each state. Estimating the immediate reward function is a simple regression problem.

The key feature of the SR representation is that, given a change in the reward function, the Q-values can be recomputed straightforwardly using (3), giving raise to immediate adaptation of the resulting policy.

2.4 Eligibility traces for the SR

As well as for model-free value-RL approaches, eligibility traces can be naturally defined for the SR approach as a matrix E of size $|\mathcal{A} \times \mathcal{S}|$ associated to a new hyper-parameter λ [37, 40]. Its update rule at each step of the episode is: $\forall s, a \in \mathcal{A} \times \mathcal{S},\ E(s, a) \leftarrow \gamma\lambda E(s, a)$ and $E(s_t, a_t) \leftarrow 1$. Given the current state occupancy TD error written as a vector:

$$\boldsymbol{\delta}_{s_t,a_t} = \begin{vmatrix} \mathbb{1}[s_t = \bar{s_1}] + \gamma M(s_{t+1}, \bar{s_1}, a_{t+1}) - M(s_t, \bar{s_1}, a_t) \\ \vdots \\ \mathbb{1}[s_t = \bar{s_N}] + \gamma M(s_{t+1}, \bar{s_N}, a_{t+1}) - M(s_t, \bar{s_N}, a_t) \end{vmatrix}$$

where $N = |\mathcal{S}|$. The update rule of M is then:

$$\forall s, \bar{s}, a \in \mathcal{S} \times \mathcal{S} \times \mathcal{A},\ M(s, \bar{s}, a) \leftarrow M(s, \bar{s}, a) + \alpha \cdot E(s, a) \cdot \boldsymbol{\delta}_{s_t,a_t}(\bar{s}).$$

3 Experiments

In this section we present the modelling work performed on three classical behavioural experiment. We intend to show that the SR approach can account for several behavioural flexibilities, we compare it with three other RL algorithmic approaches.

3.1 Latent learning (Blodgett, 1929)

The concept of latent learning was introduced by Blodgett [3]. It denotes the ability of animals to learn about the environment structure even when there is no specific goal to achieve nor rewarding events. Once a specific reward is introduced after latent learning, such a latent knowledge appears to enable the agent inferring a good policy significantly faster than without a preliminary latent learning step.

This kind of cognitive aptitude was mainly studied in the first half of the twentieth century, providing a decisive evidence to the "cognitive map" theory thoroughly defended by Tolman [47, 46, 45, 44] and others [33]. This cognitive ability appears crucial to understand the learning and decisional processes in animal and might be a good starting point to reconsider the relevance of dual RL theories [15, 6]. Thus, our first RL model comparison tackles this important issue of latent learning in order to bring decisive quantitative and statistical evidences for a unique comprehensive model.

3.1.1 Outline of the experiment

Blodgett conducted the first experiments on latent learning in rats [3]. He allowed some rats to discover a maze for a few days without rewarding them defining a latent learning period, before introducing a reinforcer at the end of the maze. Overall, he used three groups of 25 to 36 rats in a 6 units T-maze with one-way doors between the T-units (Figure 1).The first two experimental groups did not get any food at the end of the maze for respectively 2 and 6 days. At days respectively 3 and 7, a reward was introduced at the end of the maze; the following day the rats completed the maze much better. The last group was control rats, that always found food at the end of the maze since the first day. For each rat, a maximum of one error was counted each time it took the wrong way in a given T-unit.

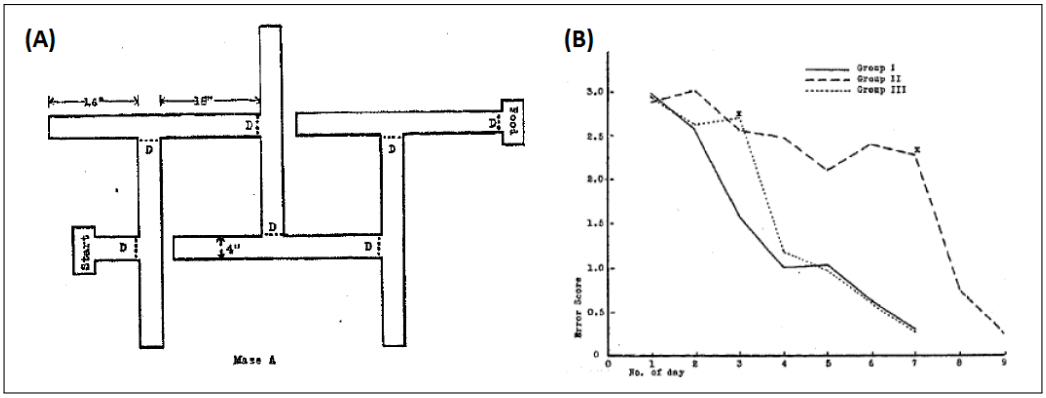

Figure 1: (A) Maze A (one-way doors are depicted by the "D" letters) and (B) results of the main latent learning experience: the curves depict the mean error score of the rats of each group (Blodgett provided no variance information), the "I" letters point the reward introduction (from Blodgett 1929).

The main results obtained by Blodgett are depicted in Figure 1. On average, the control group has a regular learning curve from the beginning of the experiment, while both other groups show a very slow decrease in their error score during the latent learning phase. The day following the reward introduction, one can notice a significant drop of the mean error score; during the following day learning continues, but on a more regular basis. This provides evidence according to Blodgett for the latent learning concept, namely the acquisition of knowledge of the environment before any reinforcement.

3.1.2 Modeling

We model the task as a deterministic Markov Decision Process (MDP) where the state space can be seen as a grid-world made of 6 T-maze as shown in Figure 2. Hence, the agent is able to go back and forth inside each T-unit as the rats sometimes do, contrary to 6 binary forced-choices MDP model designed in other similar modelling experiment [48].

In order to account for the observed mean of 3 errors at the first trial, a small bias is added to the MDP to ensure that an agent making a random walk will do 3 errors in mean at each iteration. The exact value of

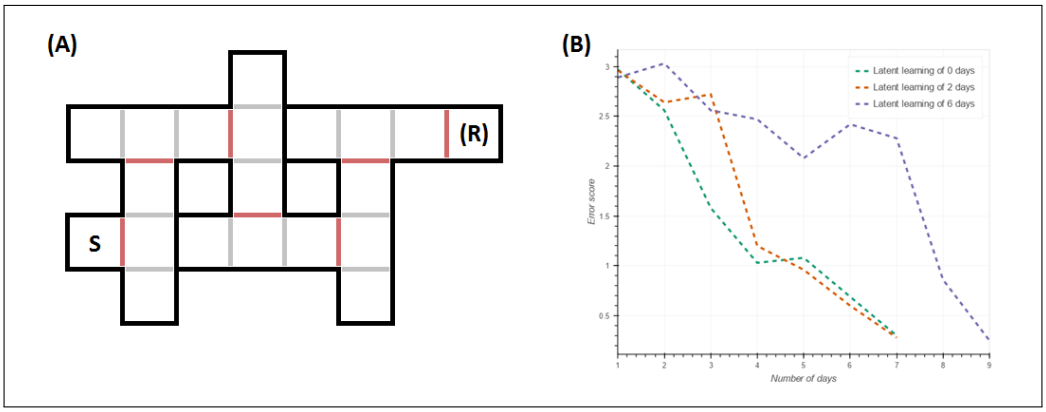

Figure 2: (A) Spatial grid-world representation of the underlying MDP. "S" is the start state, "(R)" the final and potentially rewarded state (once it is rewarded, $r = 1$), the red divisions represent the one-way doors. (B) experimental data of the Blodgett's experiment in a modern fashion (see Figure 1).

the bias is calculated given the structure of the MDP. The rationale behind this mean of 3 errors may be the ability of the rats to sometimes distinguish between dead-ends and one-way doors or their tendency to keep their initial chosen direction.

During latent learning phase, there is a slow decrease of the mean error score despite the absence of any direct reinforcement at the end of the maze. It may be explained by the fact that the rats are eventually fed outside the maze, and that this feeding phase though it occurs long after the experiment may be associated with the end of the maze. Another, but not necessarily incompatible, explanations may be that the end of the maze could be associated with an indirect reward because of the resting time it provides of finishing it, or that the rats become more and more able to distinguish between dead-ends and one-way doors. Thus, and as in [48], we introduced a supplementary hyper-parameter which specifies a tiny pseudo-reward at the end of the maze for the latent learning phase.

We define several models to investigate to what extent they are able to fit the experimental data and transcribe some of their qualitative characteristics. On the one hand, we implement an on-policy model-free RL agent supplemented with eligibility traces (Sarsa(λ)) as well as a model-based RL agent [41]. We also implement the SR approach with eligibility traces as depicted before. Finally, we implement the influential dual model initially proposed by [15] which linearly combines the Q-tables computed according to two competing model-free and model-based components.

The fitting procedure is the following: for each of the models, we proceed with a grid-search over the relevant hyper-parameters on 36 agents. With $\alpha, \lambda, \gamma, \gamma$-MB $\in [0, 0.2, 0.4, 0.6, 0.8, 1]^4$, $\beta \in [10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}, 1, 10]$, w-Hybrid $\in [0, 10^{-3}, 3.10^{-3}, 10^{-2}, 3.10^{-2}, 10^{-1}, 3.10^{-1}, 1]$ and pseudo-Reward $\in [10^{-5}, 10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}]$. Thus, each instance (each particular hyper-parameter setting) of each model is run in the modeled environment 36 times, the averaged behaviour is computed. Then, the Akaike Information Criterion (AIC) and its corrected version are used to determine the relative score of each averaged instance given the experimental baseline [30]. Finally, the best instance of each model are compared to infer the best model relatively to the set of models considered.

3.1.3 Results

The results are depicted in Figure 3. They correspond to the best instance obtained for each model given the fitting procedure presented in the previous paragraph. The final AIC scores are shown in Table 1 and the best hyper-parameters are detailed in Table 2. The model-based algorithm immediately computes an optimal policy since it discovers the position of the reward, its decision procedure is randomized to get closer to the experimental data. Nonetheless, this approach barely matches the learning behaviour of rats. Regarding the results of the model-free approach, learning is effective but that it is unable to benefit from the previous latent learning phase.

The algorithm based on the Successor Representation approach obtains the best AIC score. The dual

Table 1: Fitting measure of Blodgett's experiment. Hyper: number of hyper-parameters, LLH: log-likelihood. AIC(c): Akaike Information Criterion (and its corrected version)

Model	Hyper	LLH	AIC	AICc
Model-based	2+1	-32.85	-26.9	-25.58
Model-free(λ)	4+1	-51.46	-41.5	-37.93
Hybrid MB-MF(λ)	6+1	-58.19	-44.2	-36.72
Successor Representation(λ)	**4+1**	**-77.20**	**-67.3**	**-63.67**

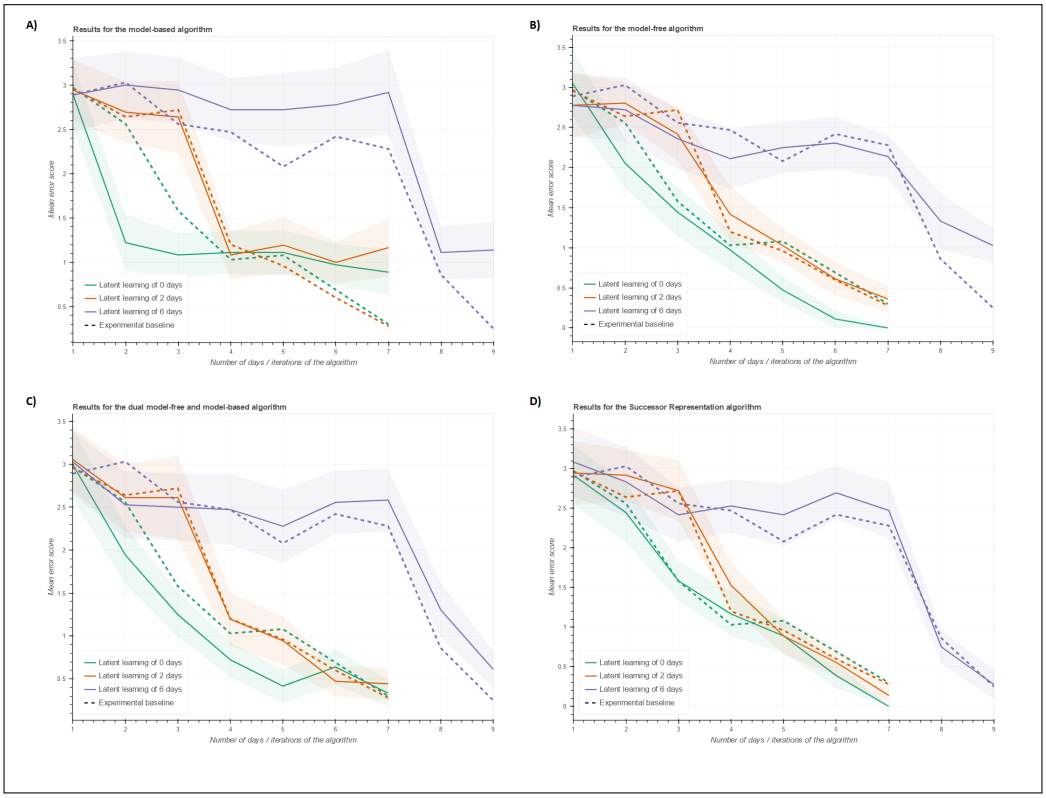

Figure 3: (A) Model-based algorithm. (B) Model-free algorithm with eligibility traces. (C) Hybrid model as in [15, 6]: a linear combination of the model-free(λ) and model-based Q-tables. (D) Successor Representation Algorithm with eligibility traces. The dotted lines represent the experimental data adapted from [3], the solid lines depict the different models (the transparent areas denote the confidence interval of 95%).

model-free and model-based algorithm seems to be far more satisfying than standalone model-free or model-based to fit the experimental data. Nonetheless, the results for the control group are poor and a more detailed analysis shows that the drops of the mean error following the end of the latent learning phase are not significantly different from the greater drop of the control group. That shows that the observed drops are not the expression of a knowledge progressively acquired during the latent learning phase but only the model-based contribution of the behaviour, which is the same, regardless of the latent learning duration.

The statistical analysis made on the top of the best models obtained following the Akaike Information Criterion is based on the non-parametric Mann-Whitney test. Its aim is to compare the distribution of the errors drop following the end of the latent learning period and the greater error drop of the control group to investigate whether they are statistically different or not. The p-values for each condition are given in Table 3.

Table 2: Best hyper-parameters for each model fitting the Blodgett's task.

Model	α	β	λ	γ	w-Hybrid	γ-MB	pseudo-Reward
Model-based		0.1			(1)	0.6	0.01
Model-free(λ)	0.6	0.001	0.2	0.6	(0)		0.0001
Hybrid MB-MF(λ)	0.4	0.001	0.4	0.4	0.003	0.2	0.001
Successor Representation(λ)	1.0	0.01	0.2	1.0			0.001

Table 3: Statistical analysis (Mann-Whitney test) of the error drop for each condition for our Blodgett's experiment modelling.

	p-value		
Model	2-latent	6-latent	Latent
Model-based	0.382	0.387	0.499
Model-free(λ)	0.479	0.155	0.267
Hybrid MB-MF(λ)	0.186	0.308	0.209
Successor Representation(λ)	0.161	**0.00518**	**0.0198**

3.2 Policy revaluation (Daw *et al*, 2011)

The purpose of Daw *et al.* was to investigate the flexibility ability of different RL models, related to observed behaviours in animals. Model-free approach alone is insufficient to account for the observed capacity of human to infer from an unexpected and indirect observation a fresh policy. In light of this, Daw *et al.* designed an experiment to highlight the complementary inputs of model-free and model-based approach [6].

The second goal of this article was to show that the model-free and model-based contribution are clearly dissociable at the neuronal level in the continuity of previous researches [35, 7, 15]. The results of this experiment was largely unexpected; indeed, the alleged contributions of each RL mechanism seem indissociable. These unexpected results opened the way to other researches that defended a more integrated but relatively puzzling vision of the dual theory [20, 12, 13, 31, 42].

Thus, we here intend to show that despite of a complex picture of multiple algorithms operating in parallel and combined in an unclear design, a unique and integrated approach is able to account for this kind of behavioural flexibility. All the more so since SR relies completely on a model-free learning principle.

3.2.1 Outline of the experiment

The behavioural part of the experiment consisted of 17 human subjects who completed a two-stage Markov decision task (Figure 4) for 201 iterations. Each of the states was comprised of two (semantically irrelevant) Tibetan characters which were associated with the two possible actions. The two first-stage actions probabilistically led to the two second-stage states, each action was associated in a privileged fashion with one the two second-stage states (70% of the time versus 30%, *common* and *rare* transitions, as denominated in [6]). These privileged associations were fixed during the experiment. The subjects were informed about the probabilistic structure and ratio of the experiment, but not about the exact mapping between states. Each of the second-stage actions were associated with a payoff probability which evolved independently at each experimental iteration, according to a slow Gaussian random walk (mean 0 and standard deviation of 0.025).

The behavioural questions underlying this experiment was about policy flexibility. For instance, whether an agent who discovers after a *rare* transition that a second-stage state is more likely of being rewarded will adjust his behaviour accordingly. It hence may be more interesting to select the first-stage action which predominantly leads to the state which was rewarded. Model-free and model-based strategies predict different policy revaluation in this case. The former should reinforce the decisional path which leads to the previously experienced reward, even if the underlying transition was unlikely. The latter should infer that the best first-stage decision to reach the rewarded second-stage state is to switch to the other first-stage action, given the higher transition probability.

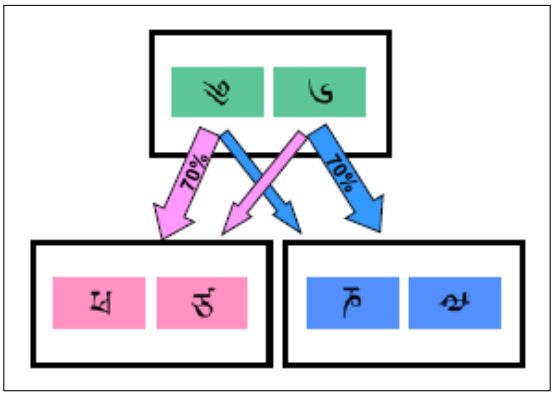

Figure 4: Experimental design of Daw *et al.* (2011). In each state of the modeled MDP (the initial state and the two final rewarded states are represented by bold rectangles) the two possible actions are depicted by the semantically relevant Tibetan characters. The transition probabilities of each first-stage action are illustrated by the arrows. The underlying reward probabilities of each second-stage actions are not shown.

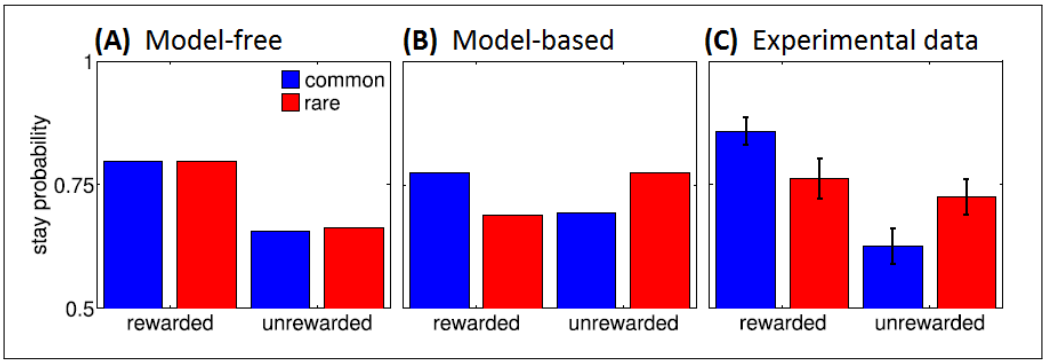

Figure 5: Daw *et al.* (2011) behavioural results (from [6]).

Figure 5 depicts the probability of repeating the previous first-stage choice given the different previously experienced patterns: a *rare* or *common* transition between the first and second-stage states leading to a *rewarded* or *not rewarded* final action. From a general perspective, with the model-free strategy, a finally rewarded choice is reinforced and is more likely to be repeated, irrespective of the probability structure of the task. Following the model-based strategy, both the transition and reward structures impact the first-stage decision in a symmetric design. The experimental data apparently suggests a combination of both properties.

3.2.2 Modeling

In their article, Daw *et al.* advocated for an hybrid algorithm combining a model-free and a model-based component to account for the patterns described in Figure 5. Thus, they adapted the dual approach of Glscher *et al.* [15] to the experimental to fit each of the participant's behaviour, adding numerous hyper-parameters particular to each stage of the Markov task. Unfortunately, they did not show the quantitative results of their hybrid approach as in Figure 5.

Table 4: Best hyper-parameters for each model fitting the Daw *et al.*'s task.

Model	α_1	α_2	β_1	β_2	λ	w-Hybrid	p	α_{reward}
Model-based		1.	1.0	0.001		(1)	1.0	
Model-free(λ)	0.2	0.4	0.0001	0.001	0.5	(0)	-3.	
Hybrid MB-MF(λ)	0.6	0.6	0.1	0.01	0.67	0.3	0	
Successor Representation(λ)	0.2	0.4	0.1	10	0.33		0	0.6

We implement the two-stage MDP as in [6], as well as several small particularities described in the Supplementary material. The MDP's structure is portrayed in the Figure 4. We thus implement the same dual RL model as in [6] (Supplementary material) (a specialized version of the one in [15]) as well as their standalone model-free and model-based versions. With the same hyper-parameters, we also implement the SR approach. The extra hyper-parameters introduced by Daw *et al.* are the following: an learning rate α for each stage, a temperature parameter β for each stage and a perseverance/switching tendency parameter p for the first-stage decision. The SR approach also needs a parameter capturing the learning rate of immediate reward value, we call it α_{reward}. In this set-up, the γ parameter is not used because it becomes redundant with the multiple α parameters. Each of the 4 models are fitted to the experimental data provided in [6] following a grid-search on 100 agents each performing 200 trials. With $\alpha_1, \alpha_2, \alpha_{reward} \in [0, 0.2, 0.4, 0.6, 0.8, 1]^3$, $\beta_1, \beta_2 \in [10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}, 1, 10]^2$, $\lambda \in [0, 0.16, 0.33, 0.5, 0, 67, 0.83, 1]$, w-Hybrid $\in [0, 10^{-3}, 3.10^{-3}, 10^{-2}, 3.10^{-2}, 10^{-1}, 3.10^{-1}, 1]$ and $p \in [-1, -0.3, -0.1, 0, 0.1, 0.3, 1]$. The fitting quality measure is the log-likelihood (LLH).

3.2.3 Results

The best fitting hyper-parameters obtained after the grid search procedure are detailed in Table 4, the resulting LLH measure in Table 5 and the best fit graphs are depicted in Figure 6. As anticipated, the model-free and model-based approach are unable to fit correctly the experimental data. Their resulting pattern are relatively similar with the anticipated one as shown in Figure 5. The dual and SR approaches fit the experiment data well, with an advantage for the SR one; interestingly, they do not require the perseveration/switching tendency parameter p to account for the experimental data.

Table 5: Fitting measure of Daw *et al.*'s experiment. Hyper: number of hyper-parameters, LLH: log-likelihood (the greater, the better).

Model	Hyper	LLH
Model-based	4	-22.06
Model-free(λ)	6	-29.13
Hybrid MB-MF(λ)	7	-10.83
Successor Representation(λ)	7	-36.57

3.3 Internal motivation (Leeper, 1935)

One of the main issue trying to model animal behaviour using RL is accounting for recurrent behaviour [41] such as the drinking/eating alternation. When you get replete but thirsty, you do not need to re-learn how to drink, and so on through an entire day. The model-free RL methods are unable to deal with this kind of natural problem, because of the entanglement between the "rewarding values of action outcomes and action outcomes *per se*" [21], that is between learning the contingencies of the environment and learning the value associated with its components.

Here we investigate whether simple and dual models relying on model-free and model-based approach can account for these daily observed capability. We intend to demonstrate that the Successor Representation approach is particularly compelling to account for multiple qualitative facets of such behaviour in addition to being computationally much less costly.

3.3.1 Outline of the experiment

The Leeper's experiment [24] is one the numerous works done on the question of the interaction between learning and internal motivation in the middle of the twentieth century [17, 24, 38, 45, 44], and considered as one of the more reliable set of results [45, 44]. It consists of 23 rats which were trained for 26 days on two mazes structurally similar to the one depicted in Figure 7. They benefited from 3 to 5 training sessions per day following a specific motivational schedule. They thus alternated between thirsty and starving conditions, respectively refereed as squares and circles in the graph of Figure 7B, that constitutes the motivational drive that eventually leads the rats to go alternately in the branch containing food or water. There was several

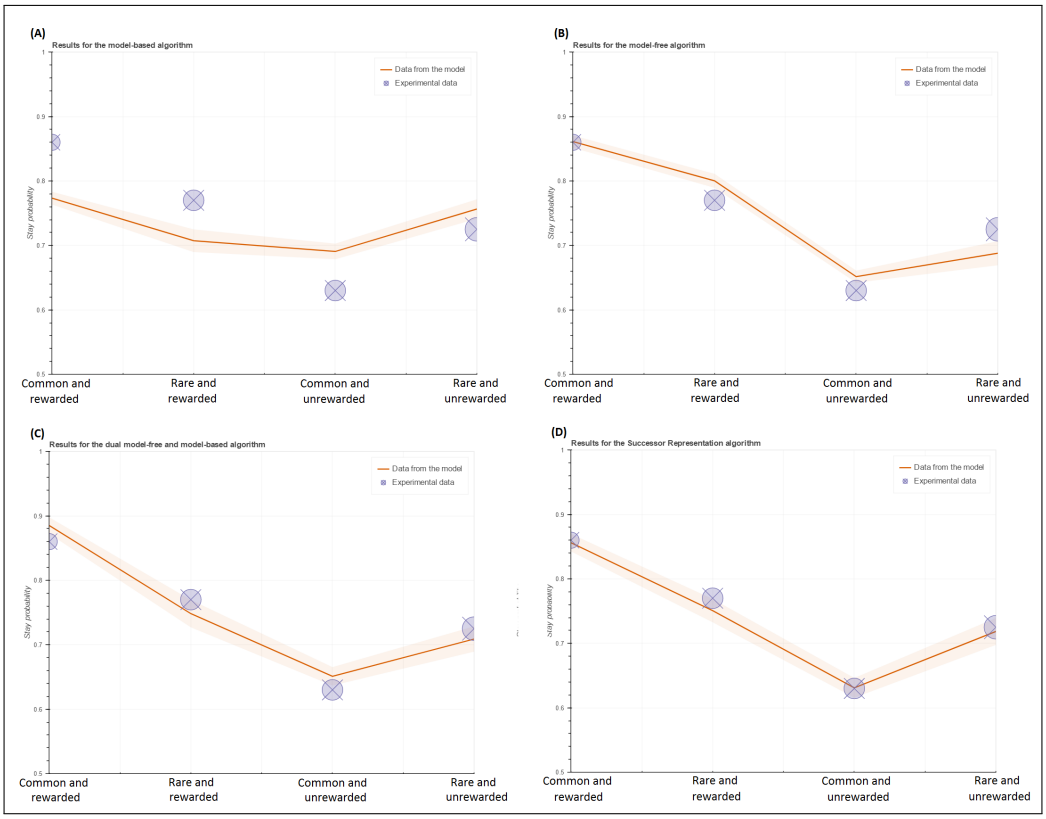

Figure 6: (A) Model-based algorithm. (B) Model-free algorithm with eligibility traces. (C) Hybrid model as in [6]: a linear combination of the model-free(λ) and model-based Q-tables. (D) Successor Representation Algorithm with eligibility traces. The crossed circle represent the experimental data adapted from [6] where the size of the circles reflect the SEM, the solid lines depict the different modelling (the transparent areas denote the confidence interval of 95%).

minor other procedural aspects. Though taking them into account in our modelling work, we do not describe them in detail here, see [24]. An error was counted if a rat went far enough to see if there was water or food at the end of the branch which did not correspond to the internal motivation of the day.

3.3.2 Modeling

We model the task as a Markov Decision Process (MDP) where the state space is seen as an Y-maze as shown in Figure 8. The final states are alternatively rewarded according to Leeper's schedule to simulate the internal motivation associated with each of them since they are respectively associated with the food and water positions.

The four models implemented are the same as in Section 3.1.2 (modeling Blodgett's experiment), except that there is here no additional hyper-parameter of pseudo-reward. As in Section 3.1.2, we proceed with a grid-search over the relevant hyper-parameters, but here on 50 agents.
With $\alpha, \lambda, \gamma, \gamma$-MB $\in [0, 0.2, 0.4, 0.6, 0.8, 1]^4$, $\beta \in [10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}, 1, 10]$, w-Hybrid $\in [0, 10^{-3}, 3.10^{-3}, 10^{-2}, 3.10^{-2}, 10^{-1}, 3.10^{-1}, 1]$ and pseudo-Reward $\in [10^{-5}, 10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}]$.
As before, we use the AIC(c) criterion to compare models.

3.3.3 Results

The results are depicted in Figure 9. They correspond to the best fitting of each model: the global AIC scores (and their corrected version) are shown in Table 7 and the best hyper-parameters are detailed in Table 6. The model-based algorithm immediately computes an optimal policy when it discovers the whole maze, its decision procedure is randomized to get closer to the experimental data: hence, there is no progressive

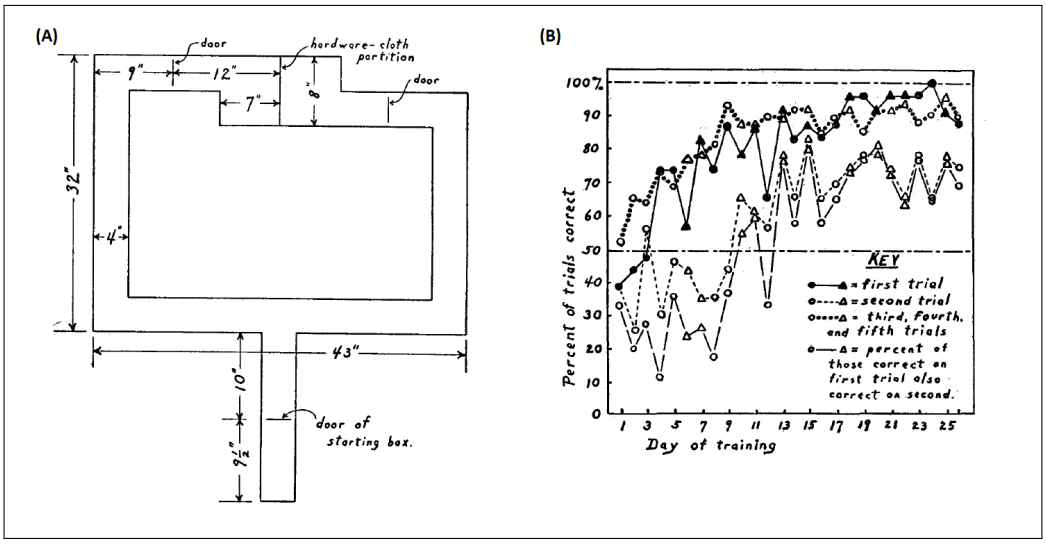

Figure 7: (A) One of the main mazes used by Leeper in his 1935 experiment and (B) the obtained results. We focus here on the solid line which is, according to Leeper himself, the main contribution of his work (from Leeper 1935).

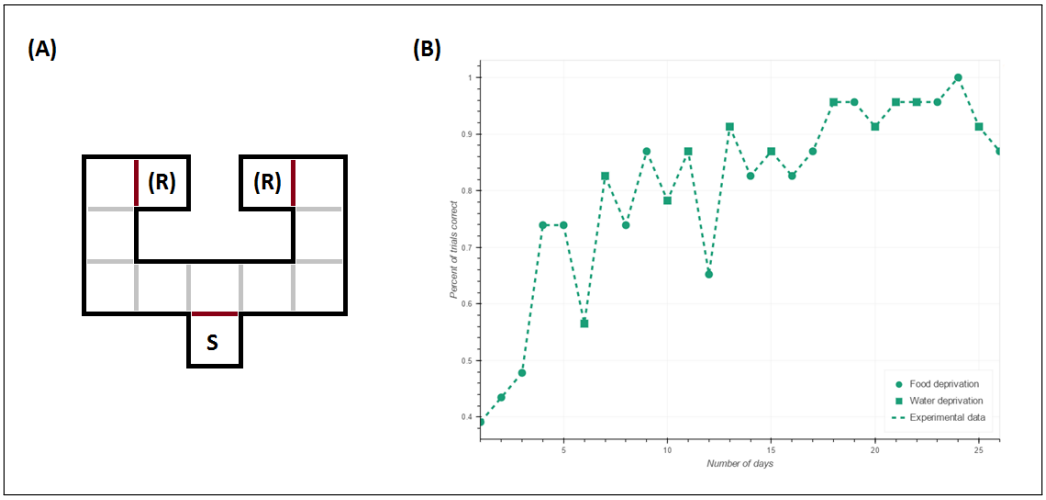

Figure 8: (A) Spatial grid-world representation of the underlying MDP. "S" is the start state, "(R)" the final and potentially rewarding states, the red divisions represent the one-way doors. (B) Experimental data of the Leeper's experiment in a modern fashion.

learning as observed in rats. The model-free approach is unable to take internal motivation into account directly. Furthermore, each time the internal motivational state changes, it needs to dismantle its previous acquired knowledge in order to laboriously re-build a new one from an adverse starting point.

The best fit of the dual algorithm is obtained with a value of 1 for the w-Hybrid parameter, which means that it relies on the model-based component only. The model-free hyper-parameters (α, λ and γ) are thus not taken into account by the algorithm. Then, it makes sense that the (A) and (C) curves of Figure 9 show the same dynamics; actually, as demonstrated in Table 6 the model-based hyper-parameters (β and γ-MB) are the same. Eventually, any model-free contribution in this dual model [15] is counter-productive in the context of this experimental setup.

Ultimately, the SR approach is the only considered model that is able to account for the behavioural flexibility expressed through these experimental data. It progressively learns to accurately choose the right branch of the maze, even though motivational drive alternates. Overall, we also observe that, as in the experimental data, there is slightly more errors the days of motivational switch than the others, which may

Table 6: Best hyper-parameters for each model fitting the Leeper's task.

Model	α	β	λ	γ	w-Hybrid	γ-MB
Model-based		0.001			(1)	0.2
Model-free(λ)	0.4	0.1	0	0.2	(0)	
Hybrid MB-MF(λ)	0.6	0.001	0.2	0.4	1	0.2
Successor Representation(λ)	0.2	0.1	0.2	0.6		

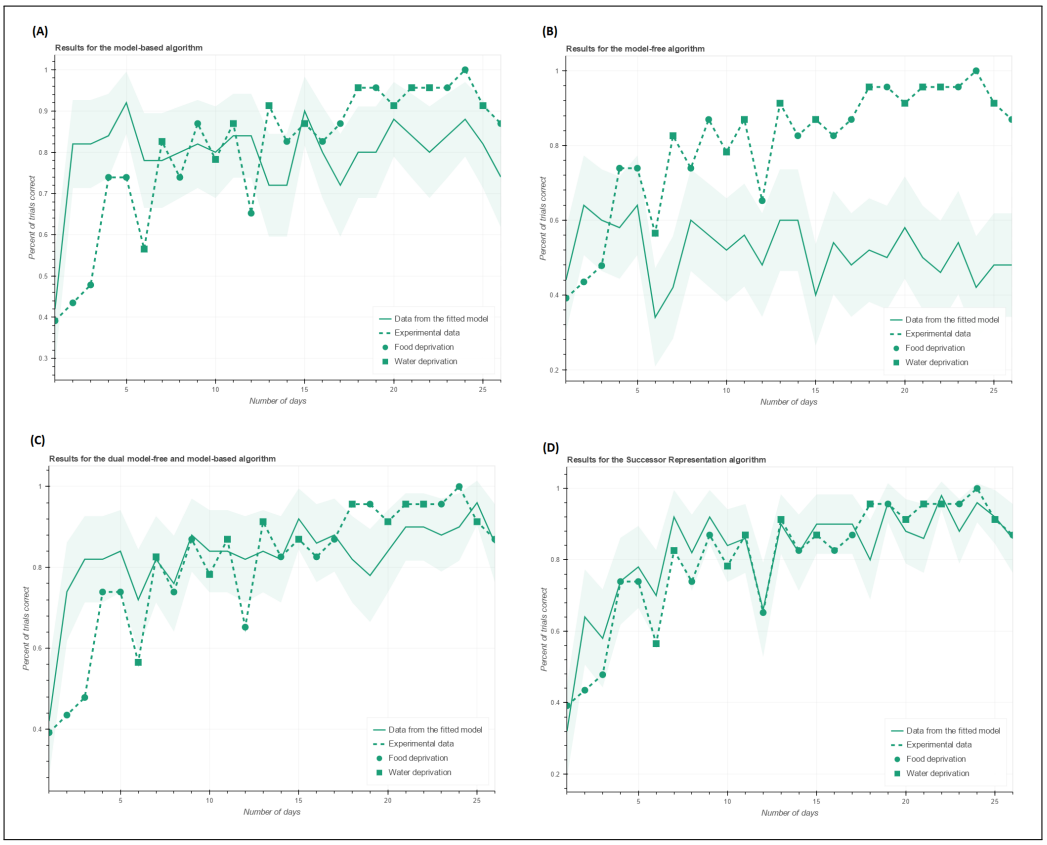

Figure 9: (A) Model-based algorithm. (B) Model-free algorithm with eligibility traces. (C) Hybrid model as in [15, 6]: a linear combination of the model-free(λ) and model-based Q-tables given the proportionality parameter w-Hybrid. (D) Successor Representation Algorithm with eligibility traces. The dotted lines represent the experimental data adapted from [24] with days of food deprivation specified with circle markers and days of water deprivation with square markers The solid lines depict the different models (the transparent areas denote the confidence interval of 95%).

be the expression of what is often refereed as model-free habitual control.

Table 7: Fitting measure of Leeper's experiment. Hyper: number of hyper-parameters, LLH: log-likelihood. AIC(c): Akaike Information Criterion (and its corrected version).

Model	Hyper	LLH	AIC	AICc
Model-based	2	-97.4	-93.4	-92.9
Model-free(λ)	4	-55.8	-47.8	-45.9
Hybrid MB-MF(λ)	6	-110.3	-98.3	-93.9
Successor Representation(λ)	**4**	**-134.2**	**-126.0**	**-124.3**

4 Related work

One of the closest works to ours is [48] in which they present their *SAwSu* algorithm. This work partly relies on the same experimental data such a [6] and [3] in order to evaluate several algorithms, but not always in the same manner as we pointed out in Section 3.1.2 and without modeling dual approaches such as [15] or [6]. Actually, the model they designed integrating associative learning and reinforcement learning may appear very similar to the SR. Indeed, it can be seen as a degeneracy of the SR model: at each time step it updates only one of the prediction component of the SR(λ) according to a delta rule instead of a TD rule. As a consequence, in a 3 states Y-MDP with two final and rewarded states in which one the final state has a greater reward although a tiny transition probability, if this unlikely state is visited just once by chance the policy will be biased toward it indefinitely. In all likelihood they did not know the RL Successor Representation introduced by Dayan in 1992 [8] as they say, speaking about their model, that "there are no other computational cognitive models that learn the spatiotemporal and the reward structures of the environment, and use both in decision-making processes".

Some other works investigate different implementations of the SR approach, trying to bridge the gap of flexibility and computational cost between classical model-free and model-based framework in a very attractive manner [29, 34]. Our work instead focuses on the standalone SR approach using TD learning which is the computationally cheaper and more biologically plausible.

5 General discussion

In this section we recall the main results of the experimental section to draw conclusions and initiate discussions about the interest of the Successor Representation for animal modelling work as well as for its appealing algorithmic flexibility.

In the latent learning experiment (Blodgett, 1929), the results show a clear superiority of the SR model, whose AIC score is smaller than the other models, regardless of the considered AIC score (classical or corrected). This superiority is confirmed by the statistical analysis inspired by Blodgett's one [3]. Hence, the SR model is the only one which shows a significant error reduction after (and only after) latent learning phase, as rats did in the Blodgett's experiment. It thus provides a decisive evidence of the ability of the SR approach to account for the latent learning phenomenon in contrast to classical and hybrid models.

One remaining question might be the meaning of the small decrease observed in rats during the latent learning phase. We hypothesized a kind of indirect or pseudo-reward inferred by the rats at the end of the maze. This might not be entirely satisfactory as the latent learning concept assumes the absence of any direct reward or goal. Nonetheless, others explanations of this decrease such as an increasing discrimination ability of the rat between one-way doors and dead-ends.

In the policy revaluation experiment (Daw *et al.*, 2011), the SR approach is the only approach which always fit the experimental data sufficiently well to always be in the confidence interval. Nevertheless, contrarily to our work on Blodgett's experiment, we found no clear evidence of a qualitative and undeniable difference between the different approaches.

Nonetheless, this approach relies on model-free learning principles such as TD learning; hence, it is almost as computationally cheap as these model-free algorithm. The SR here shows the ability to immediately infer an efficient new policy after reward variation, even when the link between the reward pattern and the subsequent optimal policy is not straightforward. Classical model-free approaches are unable to do so, it is thus typically considered as a model-based capacity [41, 29, 34].

In the internal motivation experiment (Leeper, 1935), the SR approach is the only considered models that is able to decently fit the experimental data. In addition, it shows a qualitative property similar to the one observed during the experiment that seems to go further than just appropriate fitting, appearing to express a kind of model-free habitual control on the top of its great flexibility. It is thus interesting to go back to the vision which Leeper shared in his article about learning and the decisional process, it echoes some of the structural aspects of the SR algorithm [24].

The major phenomenon put forward by Leeper [24] was the behavioural ability of the animal to progressively learn about the structure of the environment, even without reinforcer. And once rewarded, the animal uses this knowledge to immediately adopt the appropriate strategy to navigate toward the goal whose internal value is currently predominant. The primary standpoint of Leeper was to advocate for a clear distinction between *acquisition* and *utilization* of knowledge, which were generally not clearly differentiated. He particularly refereed to previous experiments that showed that "the habits or knowledge in such a case [Ed.: rats were shocked either in the correct or in the incorrect pathway] are independent of the specific motivation which was related to their development" [24]. Thus, the SR learning procedure of the *successor states* matrix could be largely independent from the reward in the environment; it remains linked with the current policy, which is in turn dependent of the rewards, but in an indirect way.

However, Leeper went one step further, defining "the phenomenon of differential motivational control of habit utilization". Which can be explained as "if motivation is importantly related to utilization, one of the prime functions of motivation would seem to be that of determining which associations, or which habits, are to be utilized in any particular situation" [24]. The parallel with the algorithmic structure of the SR approach is striking. Indeed, in SR, motivation is encoded in the reward matrix which is multiplied - at the decision (or utilization) time with the SR matrix to evaluate which action is the more relevant. Furthermore, the SR matrix encodes the state occupancy dynamics learned under the agent policy; thus, it could be viewed as a particular way to encode habits.

6 Conclusion

In this paper, we have studied the capability of the approach to account for behavioural flexibility, fitting it to experimental data published from 3 classical experiments. For each experiment we implemented 3 classical RL algorithms besides the Successor Representation. Our main concern was to compare the SR approach to the model-free and model-based dual models. Thus, we implemented the hybrid mechanism introduced by Glscher *et al.* [15] and adapted by Daw *et al.* [6].

We were aiming to show that the Successor Representation implementation can on its own account for subtle and varied behavioural flexibilities, well beyond the raw reward revaluation paradigm. The results presented here demonstrate that the RL Successor Representation approach is of special interest to account for several flexibility abilities such as a latent learning, immediate policy revaluation and internal motivation variation. Thus, the SR seems to be particularly suited to model animal behaviour. Nevertheless, an important limitation of our study is that we did not have access to individual data to get a more constrained fit of the algorithms. Another point is that we could compare the SR to several other dual mechanisms such as [23, 20, 31]. Conversely, it would be interesting to determine whether the SR can address other behavioral phenomena that have been explained by dual mechanisms, such as the "sign-tracker" "versus goal-tracker" behaviors [25] or negative automaintenance in pigeons [26]. Finally, our study leaves open the question of central interest of the actual neurophysiological implementation of the learning mechanism [14, 29, 39]. Along a different line of thought, the model-free RL framework is getting increasingly used in real-world engineering context [27, 28, 32], sometimes using the SR approach [22, 18, 2]; hence our work provides decisive clues of which flexibility capabilities the SR could provide to these implementations.

References

[1] BALLEINE B. W., DAW N. D. & ODOHERTY J. P. (2008). Multiple forms of value learning and the function of dopamine. *Neuroeconomics: decision making and the brain*, **36**, 7–385.

[2] BARRETO A., MUNOS R., SCHAUL T. & SILVER D. (2016). Successor features for transfer in reinforcement learning. *arXiv preprint arXiv:1606.05312*.

[3] BLODGETT H. C. (1929). The effect of the introduction of reward upon the maze performance of rats. *University of California publications in psychology*.

[4] COLLINS A. G. & FRANK M. J. (2016). Surprise! dopamine signals mix action, value and error. *nature neuroscience*, **19**(1), 3–5.

[5] DAW N. D. & DAYAN P. (2014). The algorithmic anatomy of model-based evaluation. *Phil. Trans. R. Soc. B*, **369**(1655), 20130478.

[6] 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.
[7] DAW N. D., NIV Y. & DAYAN P. (2005). Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioral control. *Nature neuroscience*, **8**(12), 1704–1711.
[8] DAYAN P. (1992). The convergence of td (λ) for general λ. *Machine learning*, **8**(3-4), 341–362.
[9] DAYAN P. (1993). Improving generalization for temporal difference learning: The successor representation. *Neural Computation*, **5**(4), 613–624.
[10] DAYAN P. (2002). Motivated reinforcement learning. *Advances in neural information processing systems*, **1**, 11–18.
[11] DAYAN P. & DAW N. D. (2008). Decision theory, reinforcement learning, and the brain. *Cognitive, Affective, & Behavioral Neuroscience*, **8**(4), 429–453.
[12] DESERNO L., HUYS Q. J., BOEHME R., BUCHERT R., HEINZE H.-J., GRACE A. A., DOLAN R. J., HEINZ A. & SCHLAGENHAUF F. (2015). Ventral striatal dopamine reflects behavioral and neural signatures of model-based control during sequential decision making. *Proceedings of the National Academy of Sciences*, **112**(5), 1595–1600.
[13] GERSHMAN S. J., MARKMAN A. B. & OTTO A. R. (2014). Retrospective revaluation in sequential decision making: A tale of two systems. *Journal of Experimental Psychology: General*, **143**(1), 182.
[14] GERSHMAN S. J., MOORE C. D., TODD M. T., NORMAN K. A. & SEDERBERG P. B. (2012). The successor representation and temporal context. *Neural Computation*, **24**(6), 1553–1568.
[15] GLÄSCHER J., DAW N., DAYAN P. & O'DOHERTY J. P. (2010). States versus rewards: dissociable neural prediction error signals underlying model-based and model-free reinforcement learning. *Neuron*, **66**(4), 585–595.
[16] HOUK J. C., DAVIS J. L. & BEISER D. G. (1995). *Models of information processing in the basal ganglia*. MIT press.
[17] HULL C. L. (1933). Differential habituation to internal stimuli in the albino rat. *Journal of Comparative Psychology*, **16**(2), 255.
[18] JADERBERG M., MNIH V., CZARNECKI W. M., SCHAUL T., LEIBO J. Z., SILVER D. & KAVUKCUOGLU K. (2016). Reinforcement learning with unsupervised auxiliary tasks. *arXiv preprint arXiv:1611.05397*.
[19] KARIMI Y. & BOLAND H. (1991). Experimental investigation of"latent learning"in mice. *The International Journal of Humanities*, **3**, 18–24.
[20] KERAMATI M., DEZFOULI A. & PIRAY P. (2011). Speed/accuracy trade-off between the habitual and the goal-directed processes. *PLoS Comput Biol*, **7**(5), e1002055.
[21] KOECHLIN E. (2016). Prefrontal executive function and adaptive behavior in complex environments. *Current opinion in neurobiology*, **37**, 1–6.
[22] KULKARNI T. D., SAEEDI A., GAUTAM S. & GERSHMAN S. J. (2016). Deep successor reinforcement learning. *arXiv preprint arXiv:1606.02396*.
[23] LEE S. W., SHIMOJO S. & ODOHERTY J. P. (2014). Neural computations underlying arbitration between model-based and model-free learning. *Neuron*, **81**(3), 687–699.
[24] LEEPER R. (1935). The role of motivation in learning: a study of the phenomenon of differential motivational control of the utilization of habits. *The Pedagogical Seminary and Journal of Genetic Psychology*, **46**(1), 3–40.
[25] LESAINT F., SIGAUD O., FLAGEL S. B., ROBINSON T. E. & KHAMASSI M. (2014a). 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.
[26] LESAINT F., SIGAUD O. & KHAMASSI M. (2014b). Accounting for negative automaintenance in pigeons: a dual learning systems approach and factored representations. *PloS one*, **9**(10), e111050.
[27] LILLICRAP T. P., HUNT J. J., PRITZEL A., HEESS N., EREZ T., TASSA Y., SILVER D. & WIERSTRA D. (2015). Continuous control with deep reinforcement learning. *arXiv preprint arXiv:1509.02971*.
[28] MNIH V., KAVUKCUOGLU K., SILVER D., GRAVES A., ANTONOGLOU I., WIERSTRA D. & RIEDMILLER M. (2013). Playing atari with deep reinforcement learning. *arXiv preprint arXiv:1312.5602*.
[29] MOMENNEJAD I., RUSSEK E. M., CHEONG J. H., BOTVINICK M. M., DAW N. & GERSHMAN S. J. (2016). The successor representation in human reinforcement learning. *bioRxiv*, p. 083824.
[30] MOTULSKY H. & CHRISTOPOULOS A. (2004). *Fitting models to biological data using linear and nonlinear regression: a practical guide to curve fitting*. Oxford University Press.
[31] PEZZULO G., RIGOLI F. & CHERSI F. (2013). The mixed instrumental controller: using value of information to combine habitual choice and mental simulation. *Frontiers in psychology*, **4**, 92.

[32] PRITZEL A., URIA B., SRINIVASAN S., PUIGDOMÈNECH A., VINYALS O., HASSABIS D., WIERSTRA D. & BLUNDELL C. (2017). Neural episodic control. *arXiv preprint arXiv:1703.01988*.

[33] REYNOLDS B. (1945). A repetition of the blodgett experiment on'latent learning.'. *Journal of Experimental Psychology*, **35**(6), 504.

[34] RUSSEK E. M., MOMENNEJAD I., BOTVINICK M. M., GERSHMAN S. J. & DAW N. D. (2017). Predictive representations can link model-based reinforcement learning to model-free mechanisms. *bioRxiv*, p. 083857.

[35] SCHULTZ W. (1998). Predictive reward signal of dopamine neurons. *Journal of neurophysiology*, **80**(1), 1–27.

[36] SCHULTZ W. (2007). Multiple dopamine functions at different time courses. *Annu. Rev. Neurosci.*, **30**, 259–288.

[37] SIGAUD O. & BUFFET O. (2008). Processus décisionnels de Markov en intelligence artificielle.

[38] SPENCE K. W. & LIPPITT R. (1946). An experimental test of the sign-gestalt theory of trial and error learning. *Journal of Experimental Psychology*, **36**(6), 491.

[39] STACHENFELD K. L., BOTVINICK M. & GERSHMAN S. J. (2014). Design principles of the hippocampal cognitive map. In *Advances in neural information processing systems*, p. 2528–2536.

[40] SUTTON R. S. (1988). Learning to predict by the methods of temporal differences. *Machine learning*, **3**(1), 9–44.

[41] SUTTON R. S. & BARTO A. G. (1998). *Reinforcement learning: An introduction*, volume 1. MIT press Cambridge.

[42] TARTAGLIA E. M., CLARKE A. M. & HERZOG M. H. (2017). What to choose next? a paradigm for testing human sequential decision making. *Frontiers in Psychology*, **8**.

[43] THORNDIKE E. L. (1911). *Animal intelligence: Experimental studies*. Macmillan.

[44] TOLMAN E. C. (1949). There is more than one kind of learning. *Psychological review*, **56**(3), 144.

[45] TOLMAN E. C. *et al.* (1948). Cognitive maps in rats and men.

[46] TOLMAN E. C. & HONZIK C. H. (1930a). ínsightsín rats. **4**, 215–232.

[47] TOLMAN E. C. & HONZIK C. H. (1930b). Introduction and removal of reward, and maze performance in rats. *University of California publications in psychology*.

[48] VEKSLER V. D., MYERS C. W. & GLUCK K. A. (2014). Sawsu: An integrated model of associative and reinforcement learning. *Cognitive science*, **38**(3), 580–598.

[49] WALSH M. M. & ANDERSON J. R. (2014). Navigating complex decision spaces: Problems and paradigms in sequential choice. *Psychological bulletin*, **140**(2), 466.

[50] WATKINS C. J. & DAYAN P. (1992). Q-learning. *Machine learning*, **8**(3-4), 279–292.

[51] WHITE L. M. (1995). *Temporal difference learning: Eligibility traces and the successor representation for actions*. PhD thesis, Citeseer.