An introduction to model-free and model-based reinforcement learning and their application to cognitive neuroscience

Mehdi Khamassi

mehdi.khamassi@sorbonne-universite.fr

Computational Neuroscience, Neurotechnology and Neuro-inspired Artificial Intelligence
Autumn School, Ulster University

26 October 2021

Interdisciplinary approach

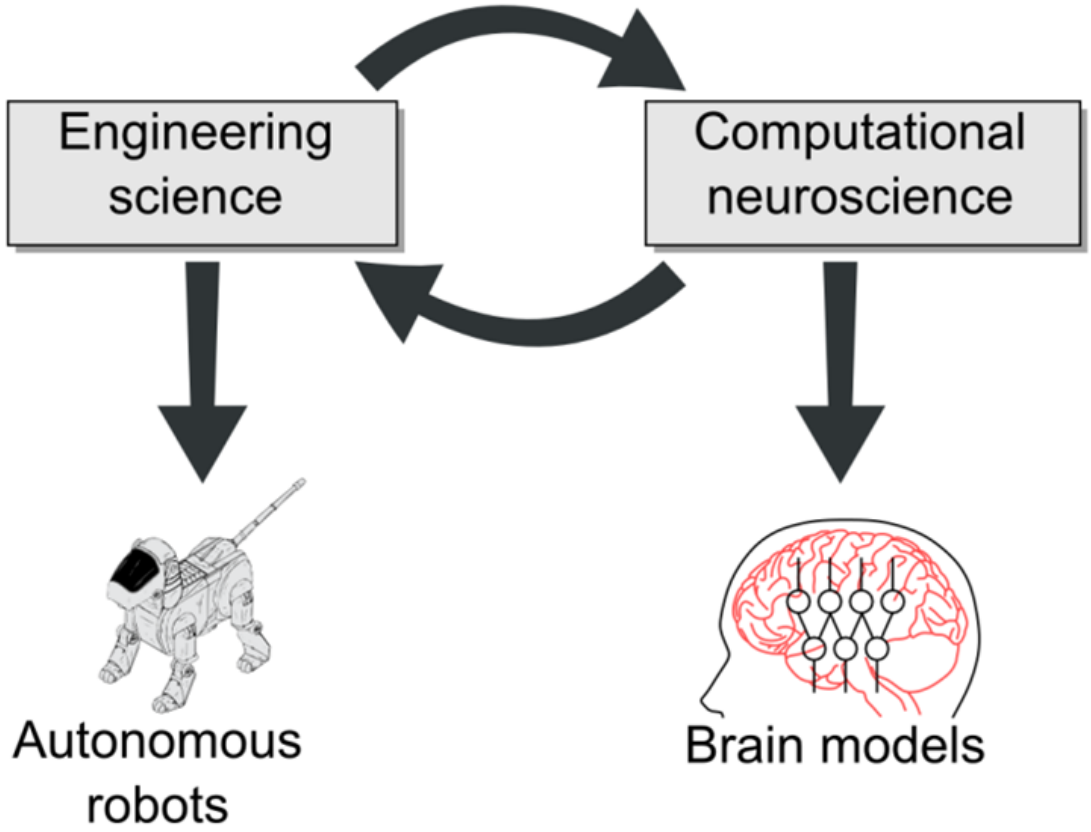


Image by Jean-Baptiste Mouret (ISIR / Sorbonne)

Multiple learning systems in the brain

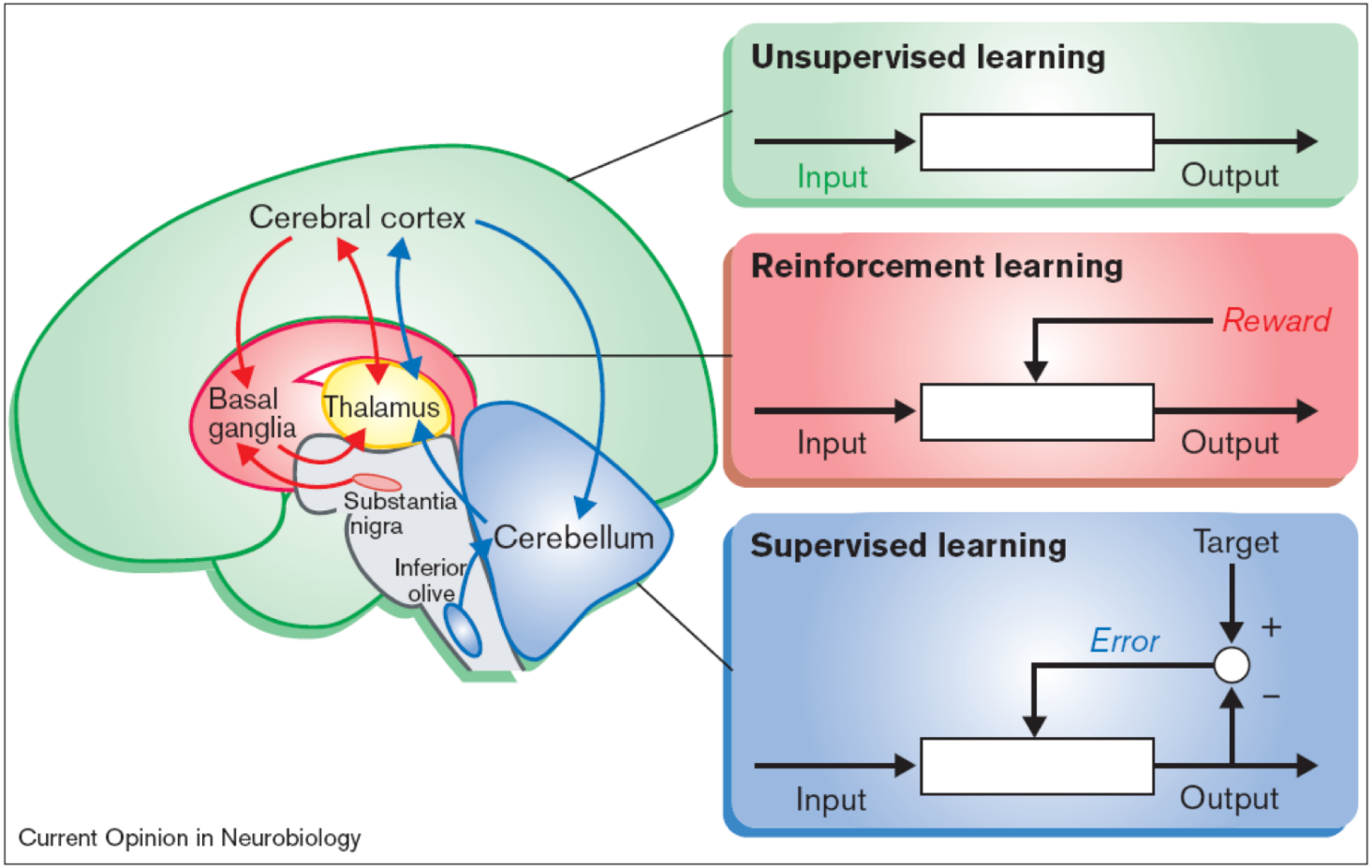
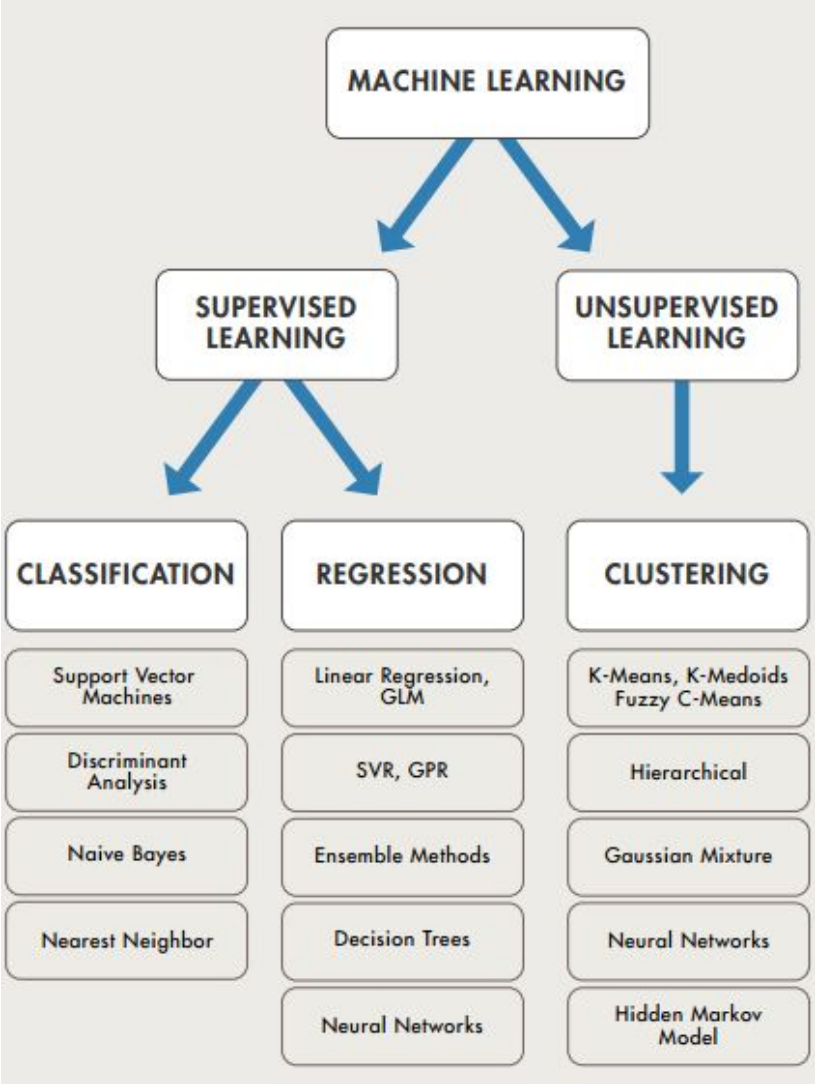


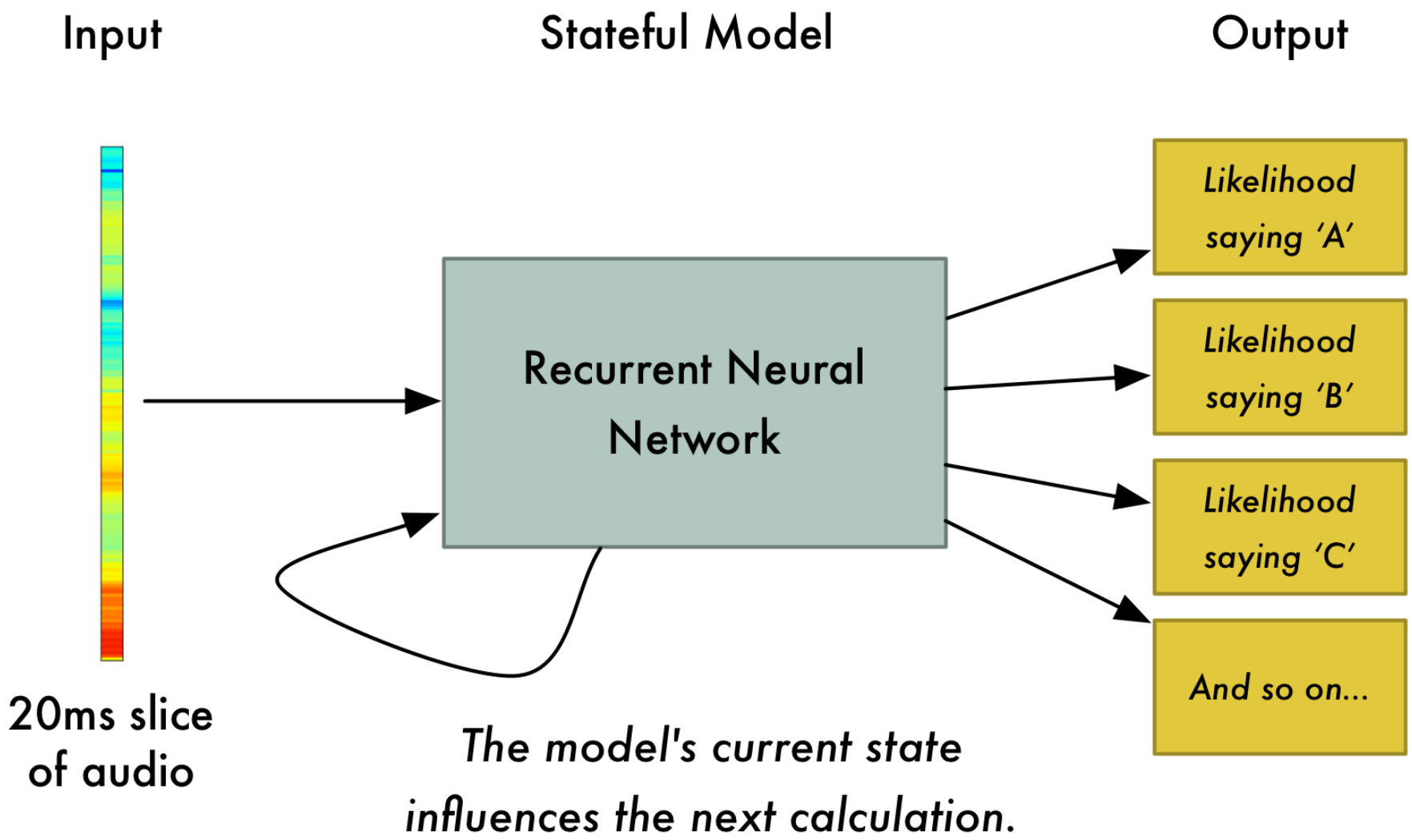
Figure by Kenji Doya (2000)

Families of learning methods

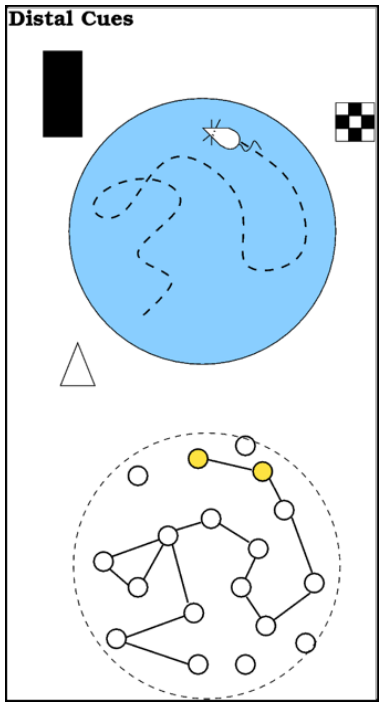


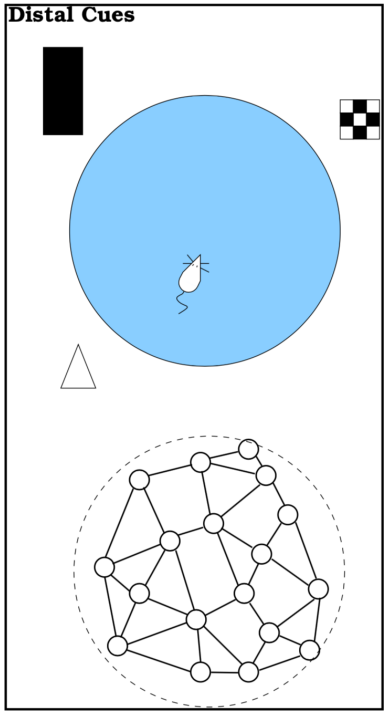
Bunker & Thabtah 2019 Applied Computing and Informatics

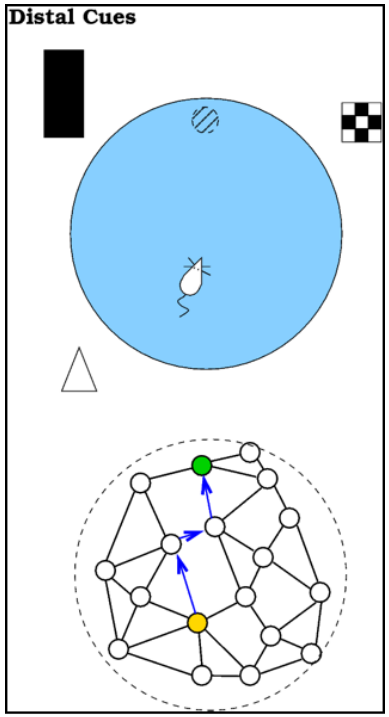
Supervised learning



Unsupervised learning





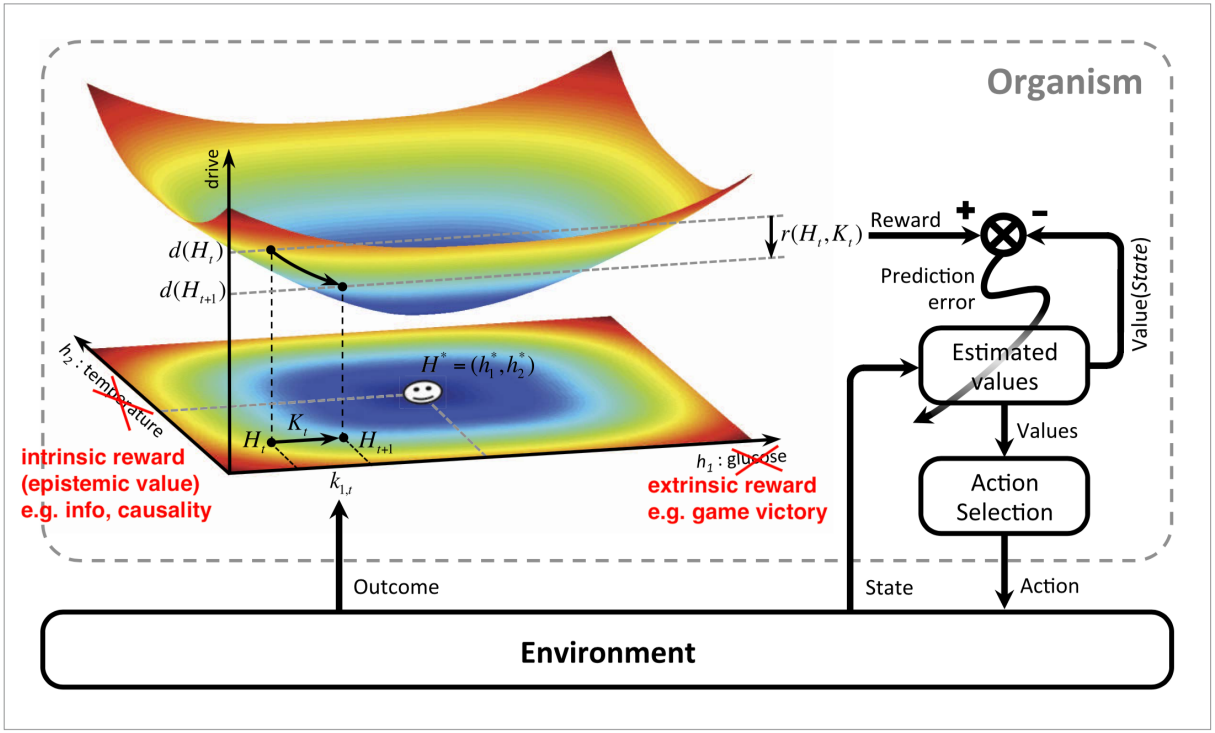


Tolman (1948) called this 'latent learning'
Figure by Benoît Girard (ISIR / Sorbonne).

Reinforcement learning

- **Decision-making:** Choice at each moment of the most appropriate behavior for an agent's survival in general, or in particular to solve a given set of tasks.
- **Reinforcement Learning (RL)** (trial/error) [Sutton & Barto 1998]: Adaptation of this choice so as to maximize a particular reward function (usually the sum of cumulative reward over an infinite horizon in a *Markov Decision Process*):
  $f(t) = \sum_{t=0}^{\infty} \gamma^t r_t$ (with $0 \leqslant \gamma \leqslant 1$).

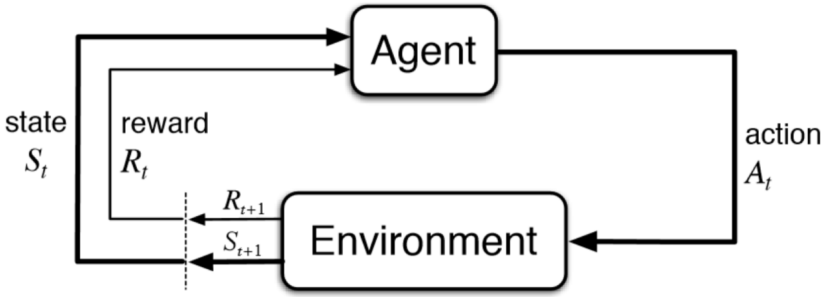
Reward function



Adapted from [Keramati & Gutkin 2014] (see also [Konidaris & Barto 2006])

- multidimensional reward functions (food, social, reproduction, information, ..)
- 'motivational' modulation of reward, e.g. through homeostatic regulation.

Markov Decision Process (MDP)



- S: state space
- A: action space
- $T : S \times A \to \Pi(S)$: transition function
- $r : S \times A \to \mathbb{R}$: reward function

- An MDP describes a problem, not a solution to that problem

[Sutton & Barto 1998]

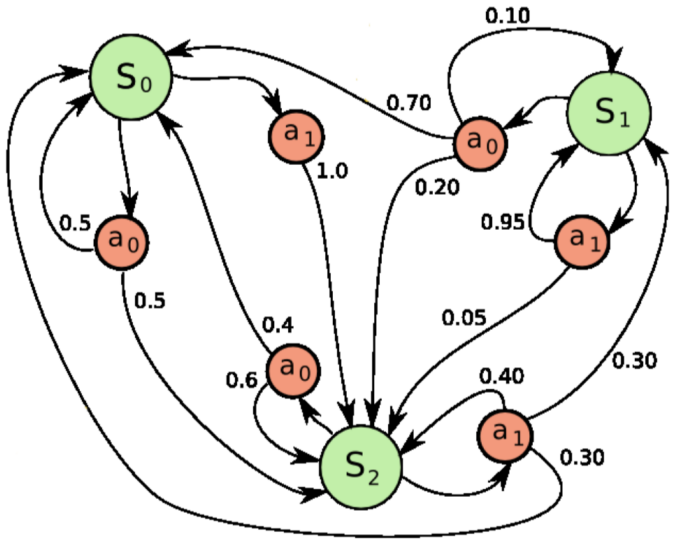
Markov Decision Process (MDP)

Markov Property:

- An MDP defines s^{t+1} and r^{t+1} as $f(s_t, a_t)$
- Markov property : $p(s^{t+1}|s^t, a^t) = p(s^{t+1}|s^t, a^t, s^{t-1}, a^{t-1}, ...s^0, a^0)$
- In an MDP, a memory of the past does not provide any useful advantage
- Reactive agents $a_{t+1} = f(s_t)$, without internal states nor memory, can be optimal

[Sutton & Barto 1998]

Example of a stochastic MDP



- Deterministic problem = special case of stochastic
- $T(s^t, a^t, s^{t+1}) = p(s'|s, a)$

Image by Olivier Sigaud (ISIR / Sorbonne)

Convention: model-based vs. model-free RL

- A **model-based (MB) agent** learns an estimate of the two functions that define a *model* of the task:
  - The reward function, $\hat{R} : (S, A) \rightarrow \mathbb{R}$.
  - The transition function, $\hat{T} : (S, A) \rightarrow \Pi(S)$.
- A **model-free (MF) agent** does not have access to this model but rather locally learns a *value function*:
  - a state value function, $V^{\pi} : S \rightarrow \mathbb{R}$ (e.g., Actor-Critic).
  - or a (state,action) value function, $Q^{\pi} : (S, A) \rightarrow \mathbb{R}$ (e.g., Q-learning).
  - or a policy function, $\pi : S \rightarrow A$ (e.g., policy search, policy gradient).

[Sutton & Barto 1998]

Navigation task from Biology: The Morris water maze

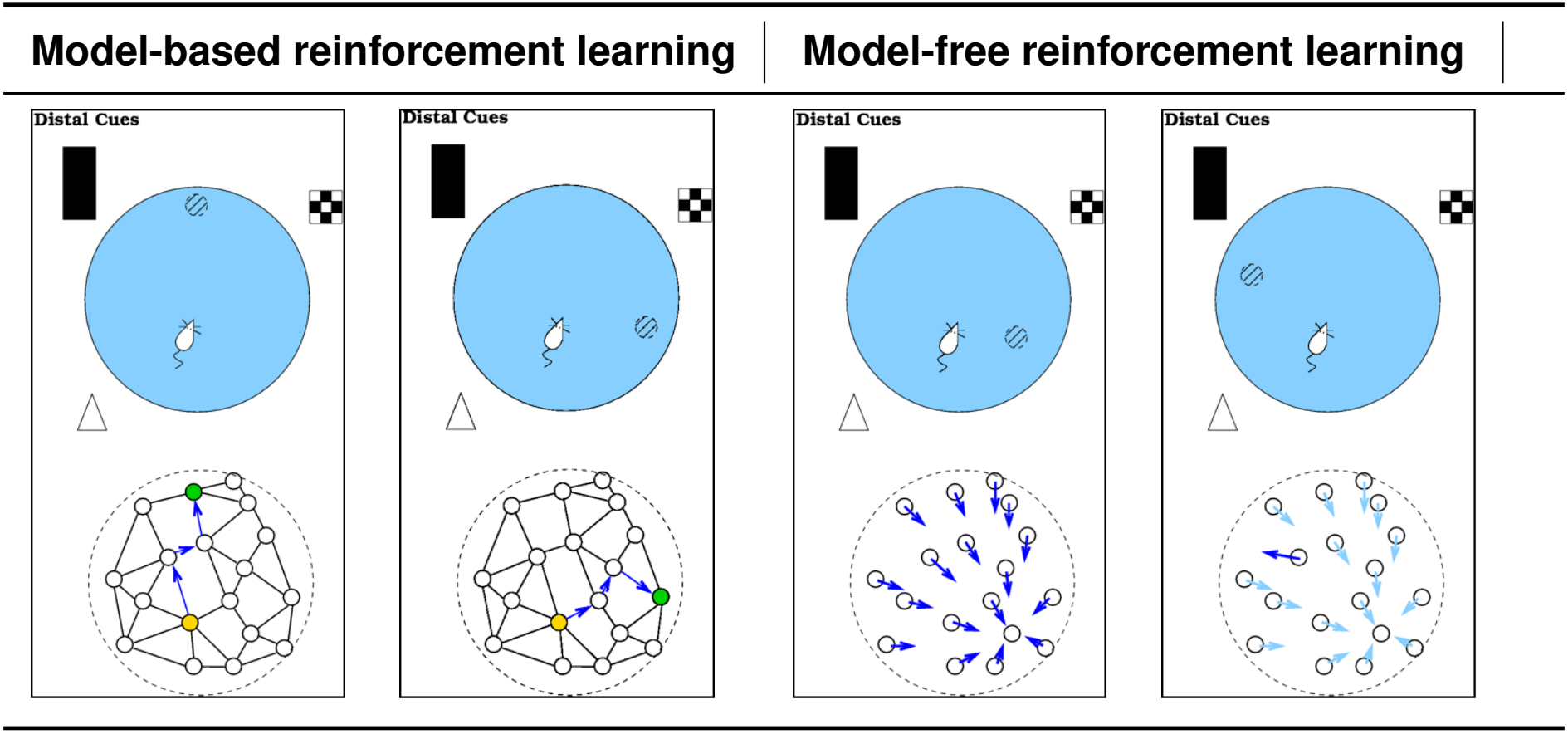


Figure by Benoît Girard (ISIR / Sorbonne). See [Khamassi et al. 2012] for a review. Also see alternative proposals, such as [Dezfouli & Balleine 2012], [Miller et al. 2019].

Take-home messages

Biology/Psychology

- Mammals' behavior typically alternates between MB and MF RL.
- Their brain includes both MB and MF RL mechanisms.

Robotics / Artificial Intelligence (AI)

- Engineering approaches to Robotics/AI typically search for an optimal solution specific to each encountered task.
- MB and MF RL turn out to be appropriate for different types of tasks [Kober et al. 2013]

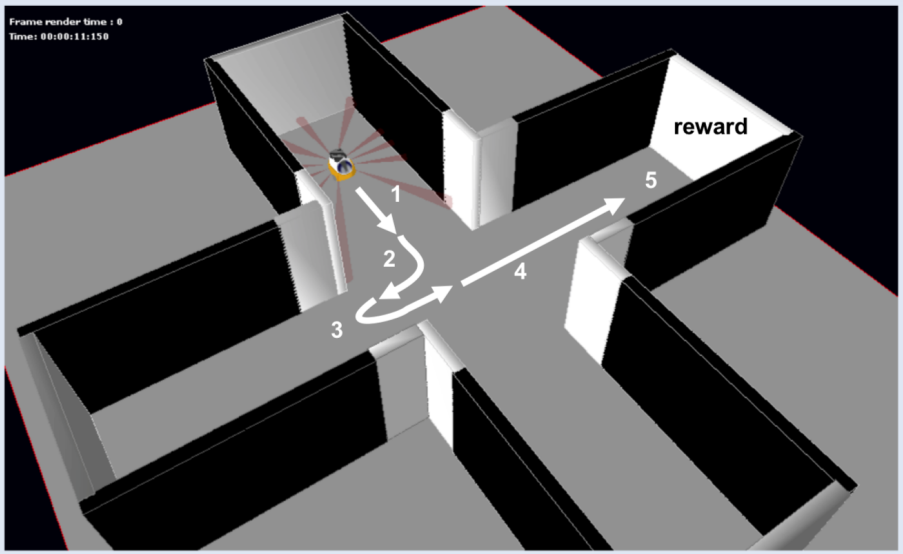
A Neuro-robotics strategy

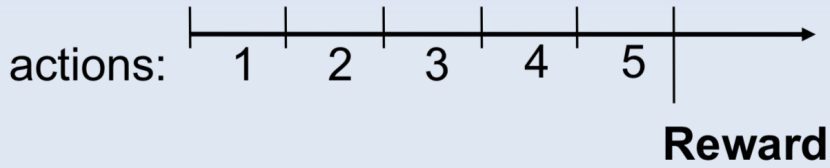
- Conceiving computational neuroscience models for the online adaptive coordination of MB and MF RL.
- Testing and improving the robustness of these models in real robots.
- Raising new biological hypotheses.

Model-free reinforcement learning

Temporal-difference learning

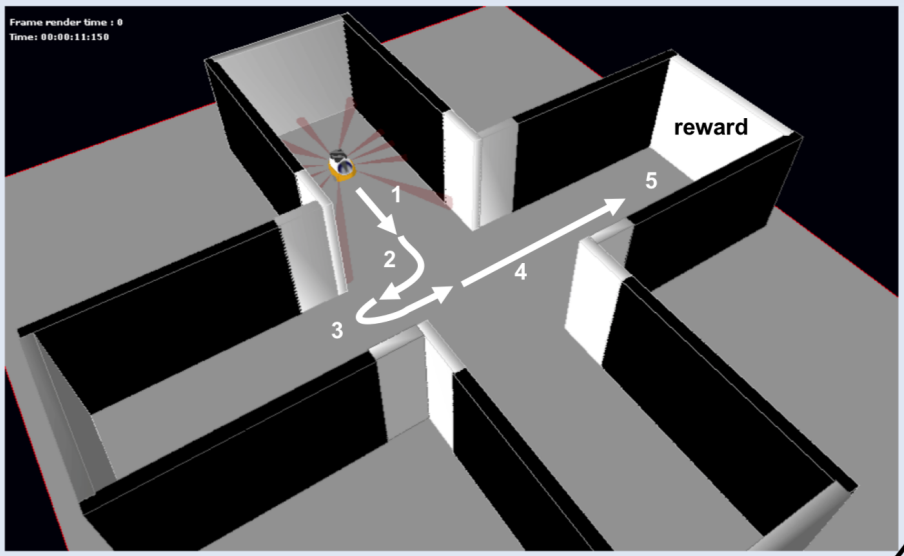
- Learning from delayed reward

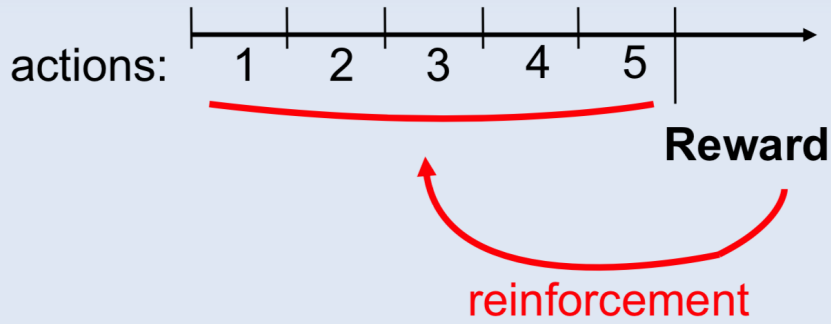




Temporal-difference learning

- Learning from delayed reward





$$\delta_t = r_t$$

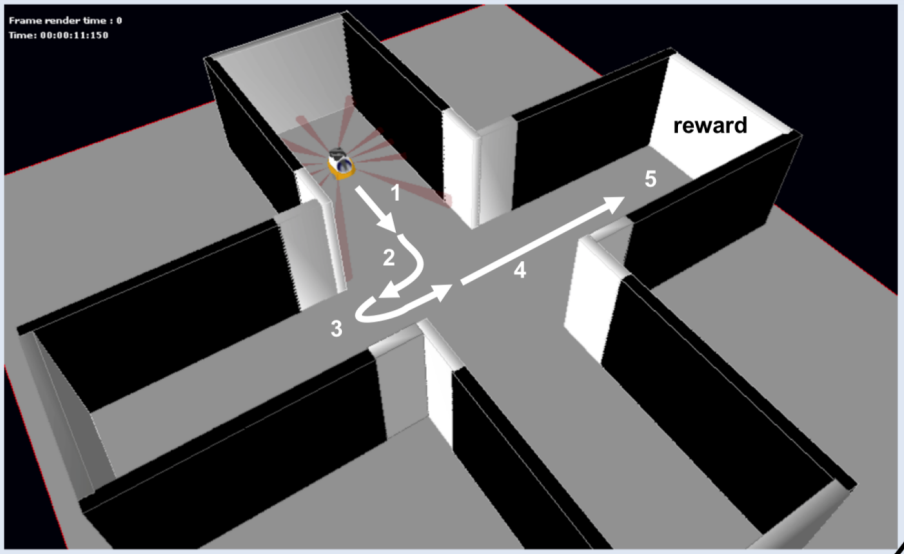
reinforcement

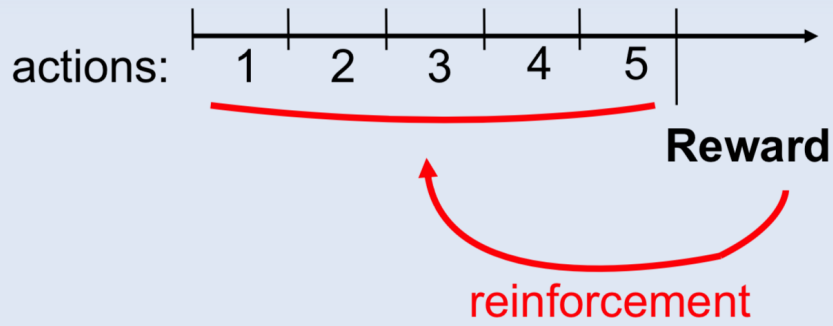
reward

Temporal-difference learning

- Learning from delayed reward

Value estimation ("reward prediction"): $\mathbf{V(s_t)}$





$$\delta_{t+n} = \mathbf{r}_{t+n} - \mathbf{V(s_t)}$$

reinforcement

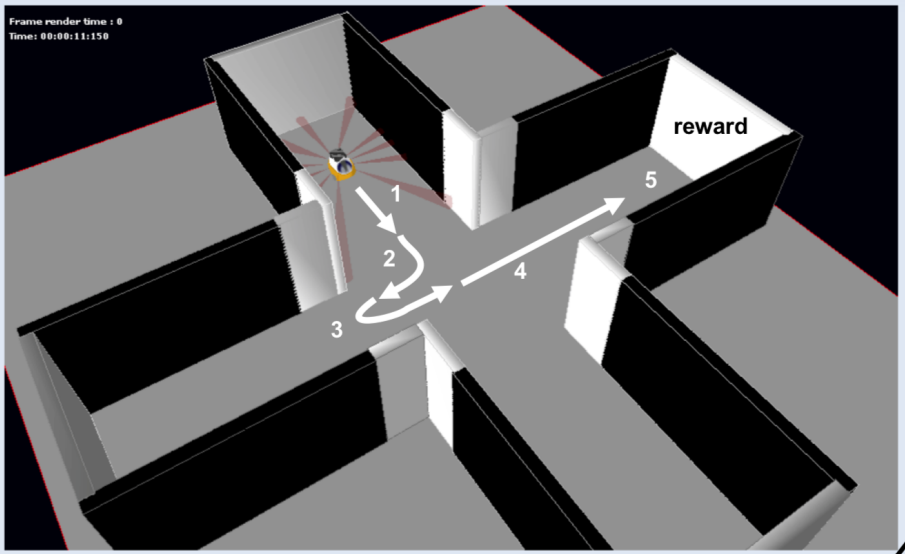
reward

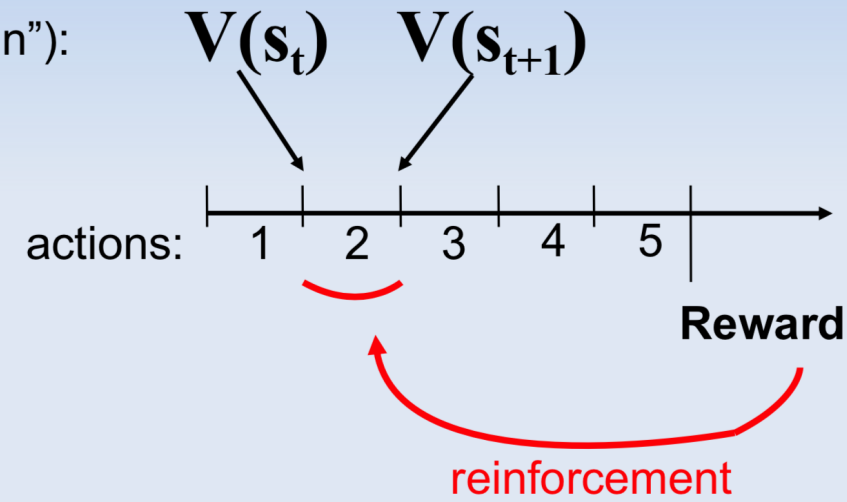
Rescorla and Wagner (1972).

Temporal-difference learning

- Temporal-Difference (TD) learning

Value estimation ("reward prediction"):





$$\delta_{t+1} = r_{t+1} + \gamma \cdot V(s_{t+1}) - V(s_t) \quad (\gamma < 1)$$

reinforcement

reward

Sutton and Barto (1998).

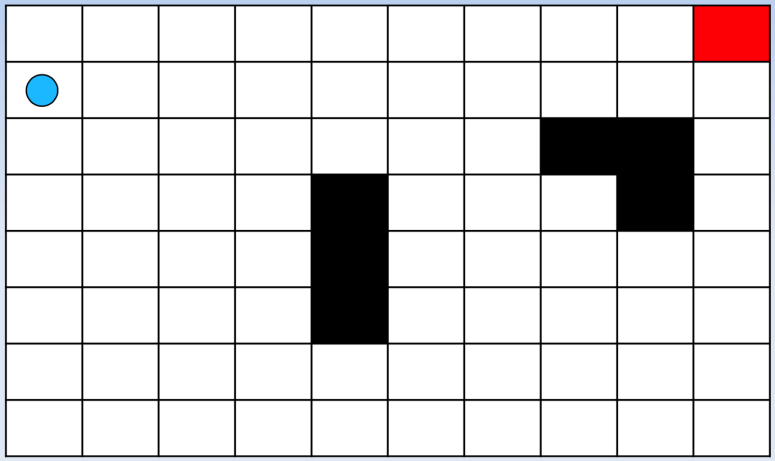
Temporal-difference learning

Learning by the method of temporal-difference:

- At each timestep, a **reward prediction error** δ_t can be computed after observing the reward r_t and the state s_t in the environment, and comparing two consecutive estimations of the value function $V(s)$.
- V-learning:
  - $\delta_t = r_t + \gamma V(s_t) - V(s_{t-1})$
  - $V(s_{t-1}) = V(s_{t-1}) + \alpha \delta_t$

[Sutton & Barto 1998]

Temporal-difference learning



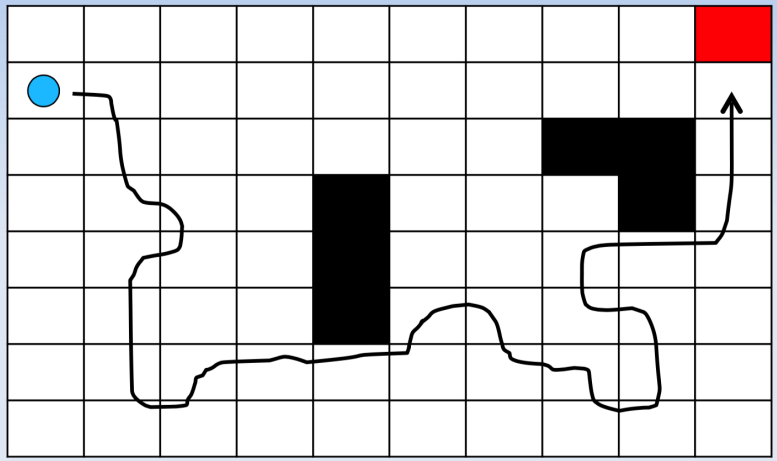
$$\delta_{t+1} = r_{t+1} + \gamma \,.\, V(s_{t+1}) - V(s_t)$$

discount factor (=0.9)

$$V(s_t) = V(s_t) + \alpha \,.\, \delta_{t+1}$$

learning rate (=0.9)

Temporal-difference learning



$$\delta_{t+1} = r_{t+1} + \gamma \cdot V(s_{t+1}) - V(s_t)$$

$$0 = 0 + 0 - 0$$

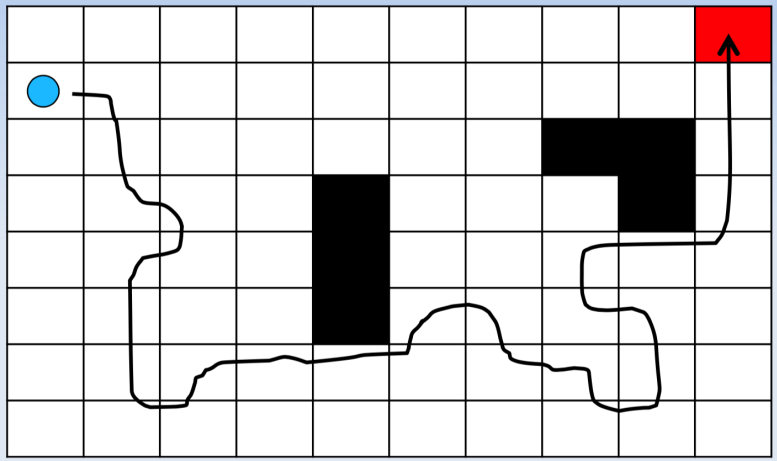
discount factor (=0.9)

$$V(s_t) = V(s_t) + \alpha \cdot \delta_{t+1}$$

$$0 = 0 + 0.9 * 0$$

learning rate (=0.9)

Temporal-difference learning



$$\delta_{t+1} = r_{t+1} + \gamma \cdot V(s_{t+1}) - V(s_t)$$

1 = 1 + 0 - 0

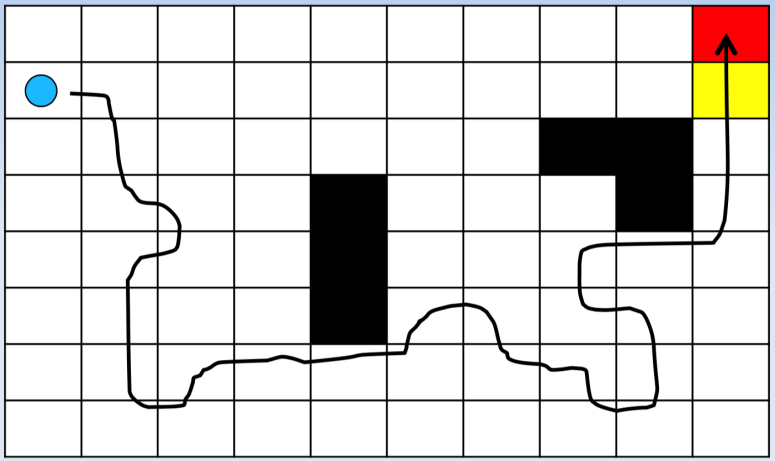
$$V(s_t) = V(s_t) + \alpha \cdot \delta_{t+1}$$

0.9 = 0 + 0.9 * 1

discount factor (=0.9)

learning rate (=0.9)

Temporal-difference learning



$$\delta_{t+1} = r_{t+1} + \gamma \cdot V(s_{t+1}) - V(s_t)$$

$$1 = 1 + 0 - 0$$

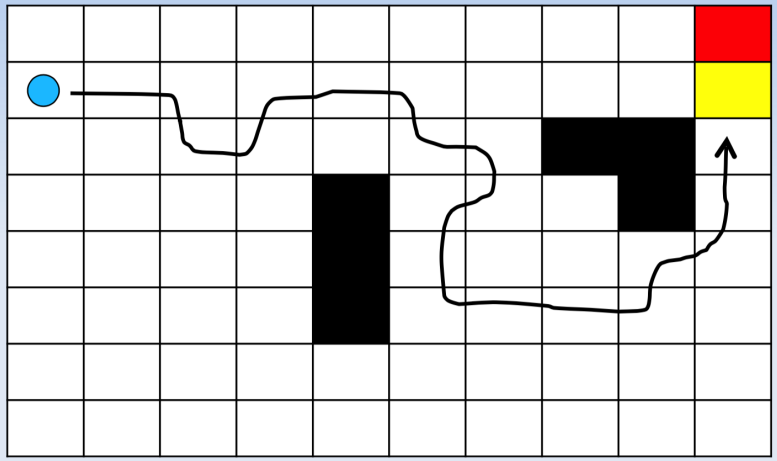
$$V(s_t) = V(s_t) + \alpha \cdot \delta_{t+1}$$

$$0.9 = 0 + 0.9 * 1$$

discount factor (=0.9)

learning rate (=0.9)

Temporal-difference learning



$$\delta_{t+1} = r_{t+1} + \gamma \cdot V(s_{t+1}) - V(s_t)$$

$$0 = 0 + 0 - 0$$

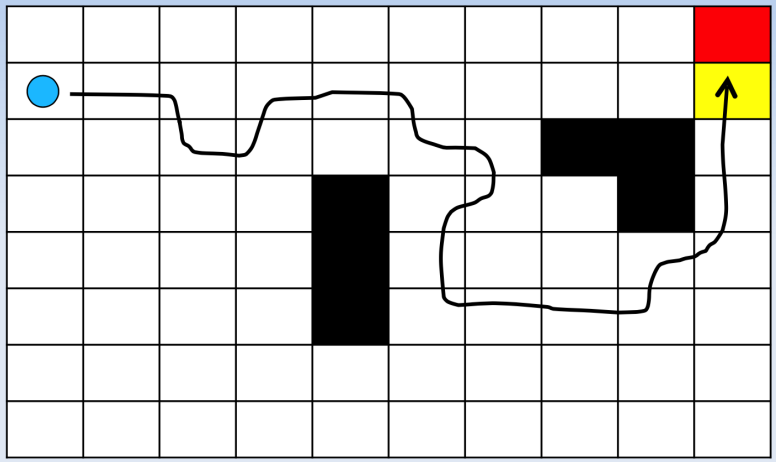
discount factor (=0.9)

$$V(s_t) = V(s_t) + \alpha \cdot \delta_{t+1}$$

$$0 = 0 + 0.9 * 0$$

learning rate (=0.9)

Temporal-difference learning



$0.81 = 0 + 0.9 * 0.9 - 0$

$\delta_{t+1} = r_{t+1} + \gamma \,.\, V(s_{t+1}) - V(s_t)$

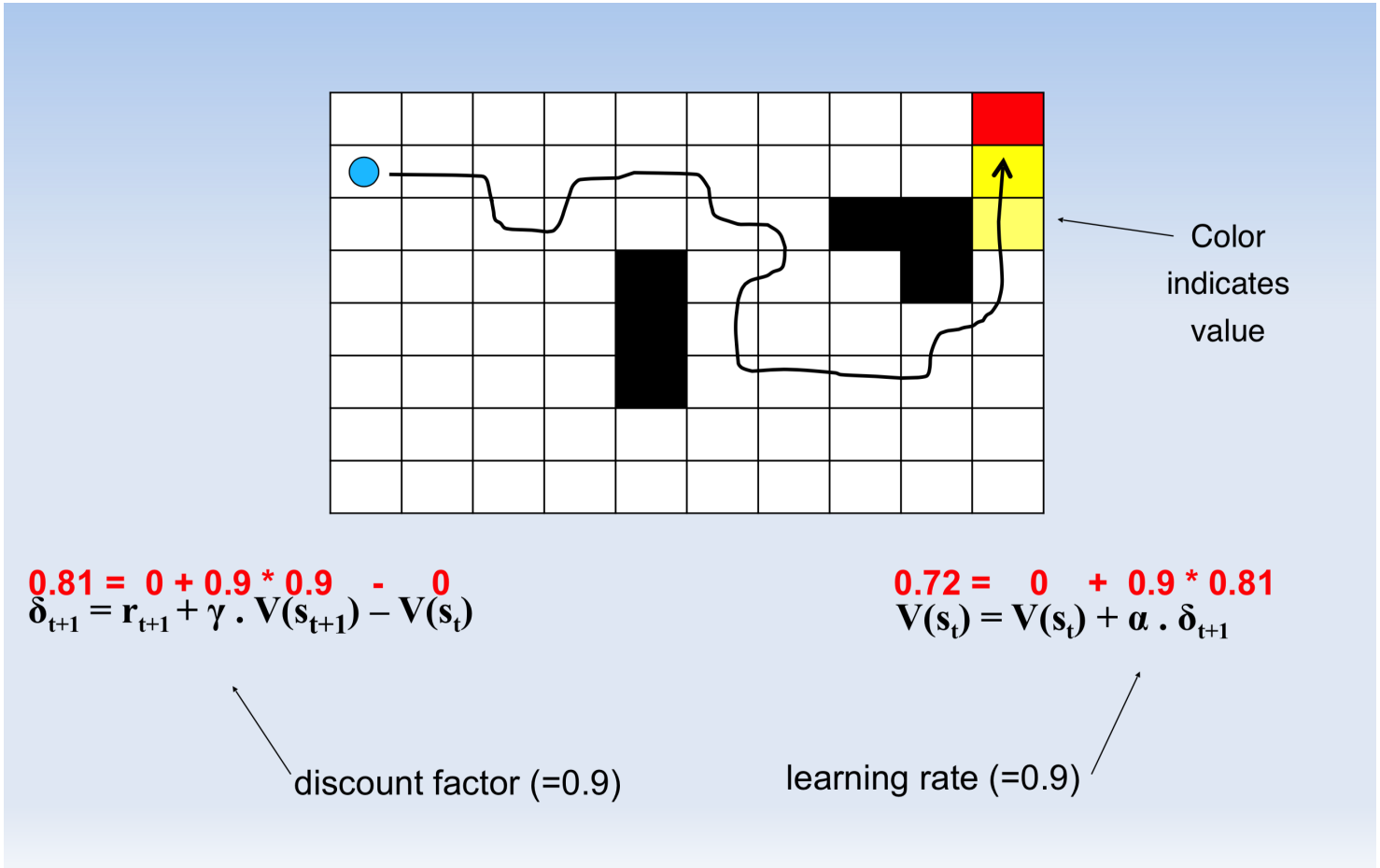
discount factor (=0.9)

$0.72 = 0 + 0.9 * 0.81$

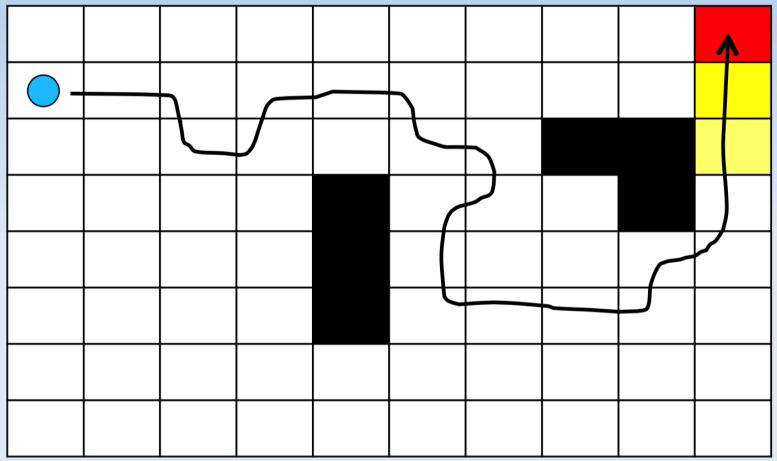
$V(s_t) = V(s_t) + \alpha \,.\, \delta_{t+1}$

learning rate (=0.9)

Temporal-difference learning



Temporal-difference learning



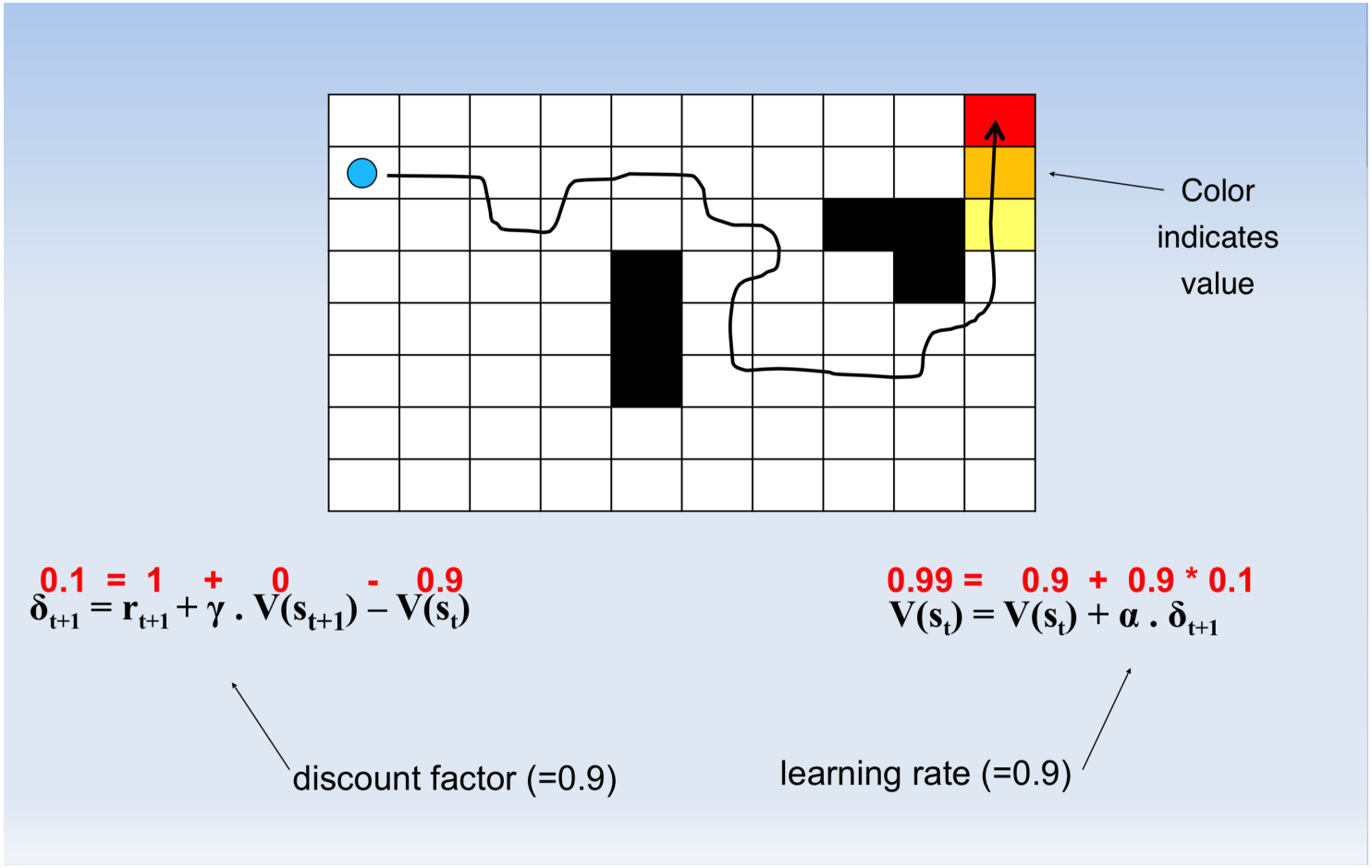
$$0.1 = 1 + 0 - 0.9$$
$$\delta_{t+1} = r_{t+1} + \gamma \,.\, V(s_{t+1}) - V(s_t)$$

discount factor (=0.9)

$$0.99 = 0.9 + 0.9 * 0.1$$
$$V(s_t) = V(s_t) + \alpha \,.\, \delta_{t+1}$$

learning rate (=0.9)

Temporal-difference learning

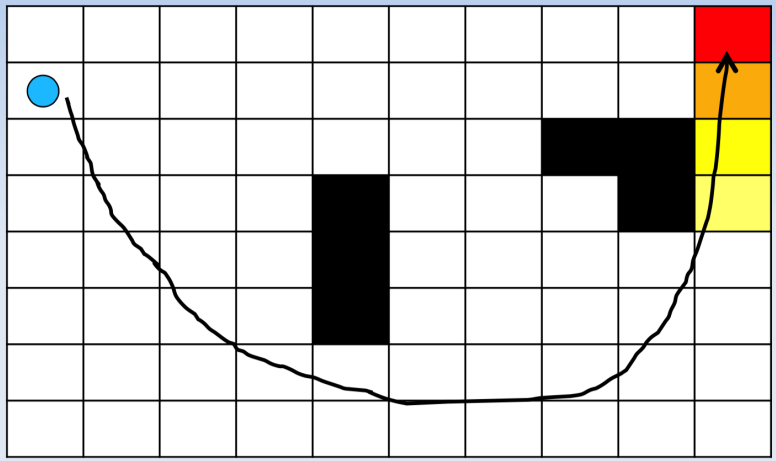


Temporal-difference learning

After

N simulations

Very long!



$$\delta_{t+1} = r_{t+1} + \gamma \,.\, V(s_{t+1}) - V(s_t)$$

discount factor (=0.9)

$$V(s_t) = V(s_t) + \alpha \,.\, \delta_{t+1}$$

learning rate (=0.1)

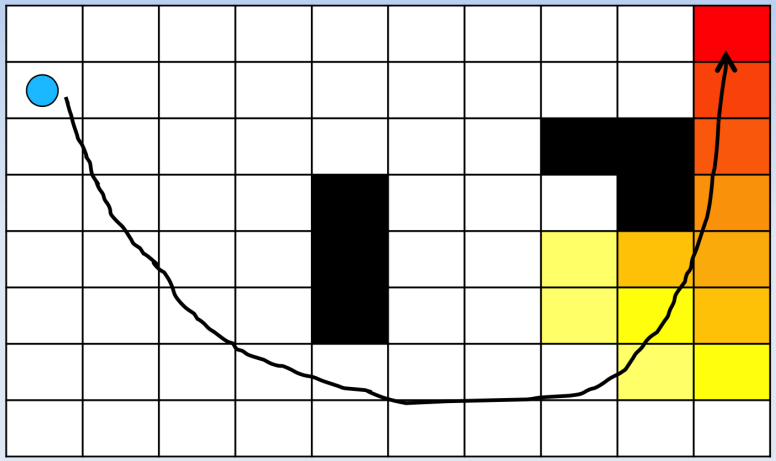
usually small for stability

Temporal-difference learning

After

N simulations

Very long!



$$\delta_{t+1} = r_{t+1} + \gamma \,.\, V(s_{t+1}) - V(s_t)$$

discount factor (=0.9)

$$V(s_t) = V(s_t) + \alpha \,.\, \delta_{t+1}$$

learning rate (=0.1)

Temporal-difference learning

After

N simulations

Very long!

$$\delta_{t+1} = r_{t+1} + \gamma \,.\, V(s_{t+1}) - V(s_t)$$

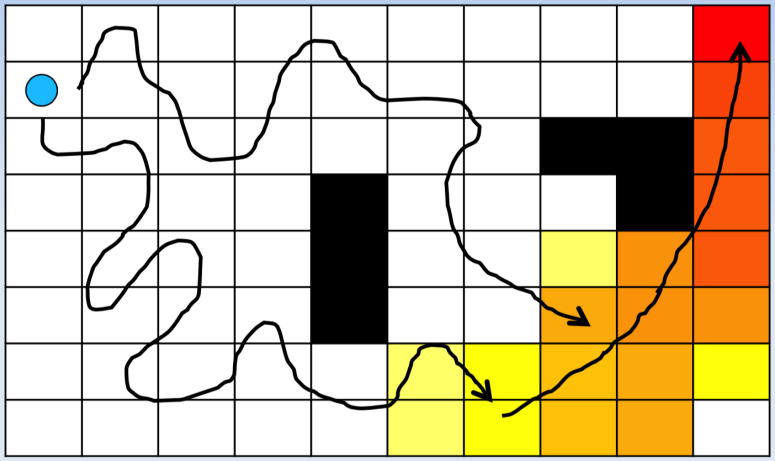
discount factor (=0.9)

$$V(s_t) = V(s_t) + \alpha \,.\, \delta_{t+1}$$

learning rate (=0.1)

Temporal-difference learning

May converge to a sub-optimal solution!



$$\delta_{t+1} = r_{t+1} + \gamma \,.\, V(s_{t+1}) - V(s_t)$$

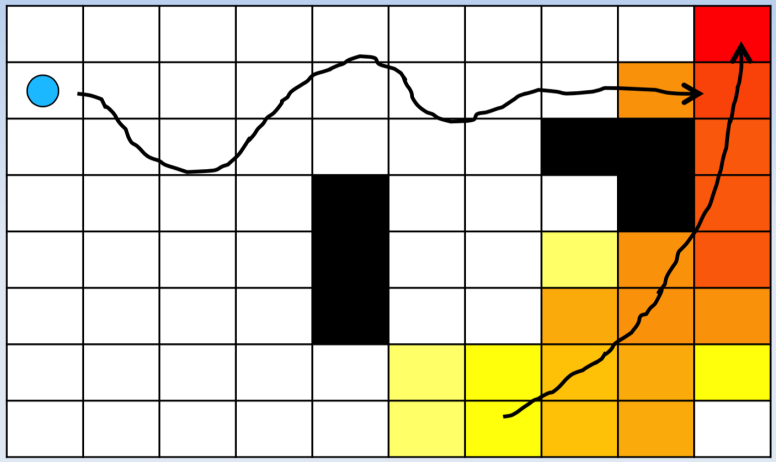
discount factor (=0.9)

$$V(s_t) = V(s_t) + \alpha \,.\, \delta_{t+1}$$

learning rate (=0.1)

Temporal-difference learning

Exploration-Exploitation trade-off



$$\delta_{t+1} = r_{t+1} + \gamma \,.\, V(s_{t+1}) - V(s_t)$$

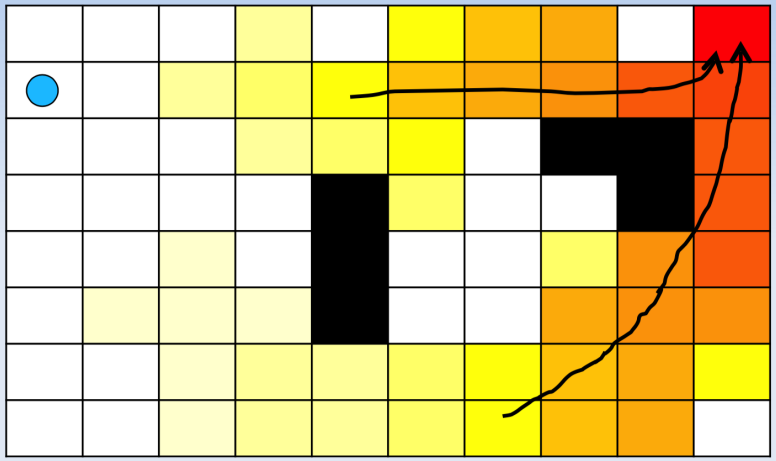
discount factor (=0.9)

$$V(s_t) = V(s_t) + \alpha \,.\, \delta_{t+1}$$

learning rate (=0.1)

Temporal-difference learning

Finds best solution after infinite time!



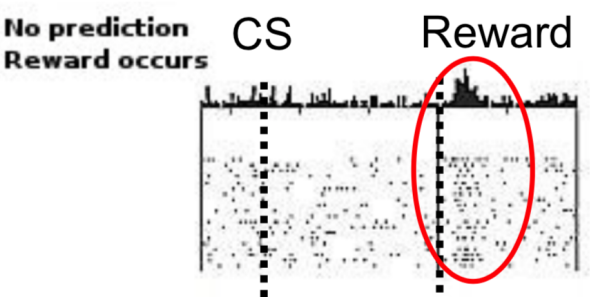
$$\delta_{t+1} = r_{t+1} + \gamma \,.\, V(s_{t+1}) - V(s_t)$$

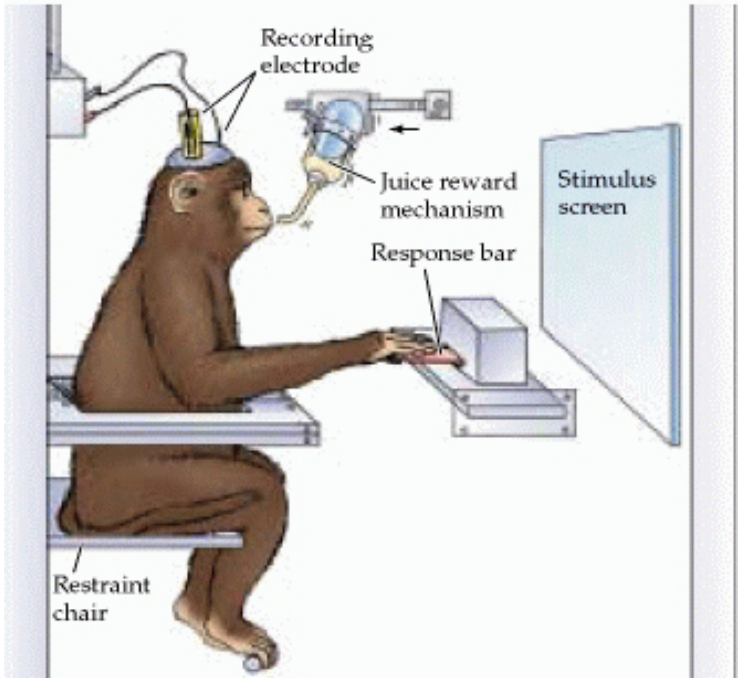
discount factor (=0.9)

$$V(s_t) = V(s_t) + \alpha \,.\, \delta_{t+1}$$

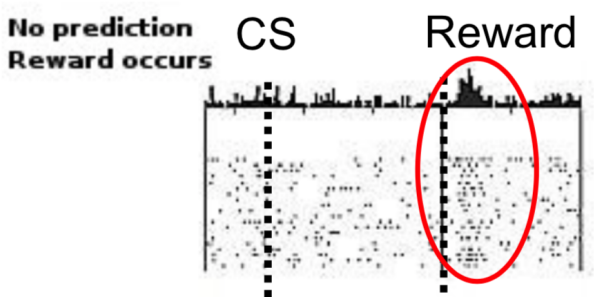
learning rate (=0.1)

Dopamine neurons' reward prediction error signal



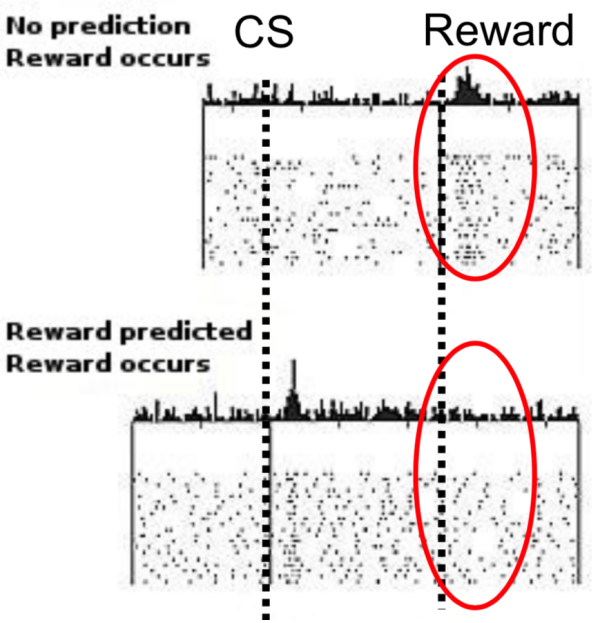


Dopamine neurons' reward prediction error signal



$$\overset{+1}{\frown}\delta_{t+1} = r_{t+1} + \gamma \cdot V(s_{t+1}) - V(s_t)$$

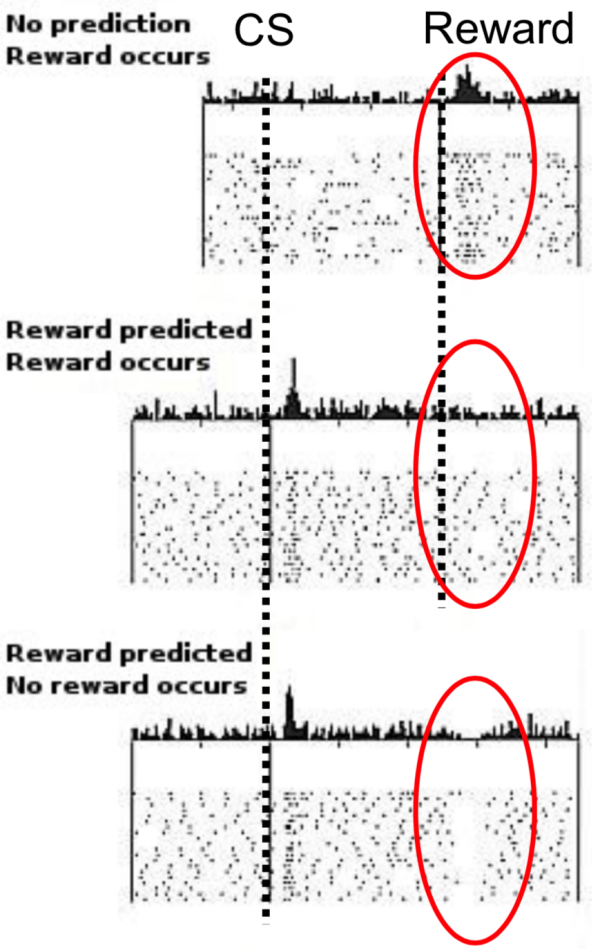
Dopamine neurons' reward prediction error signal



$$\overbrace{\delta_{t+1}}^{+1} = r_{t+1} + \gamma \cdot V(s_{t+1}) - V(s_t)$$

$$\overbrace{\delta_{t+1}}^{0} = r_{t+1} + \gamma \cdot V(s_{t+1}) - V(s_t)$$

Dopamine neurons' reward prediction error signal



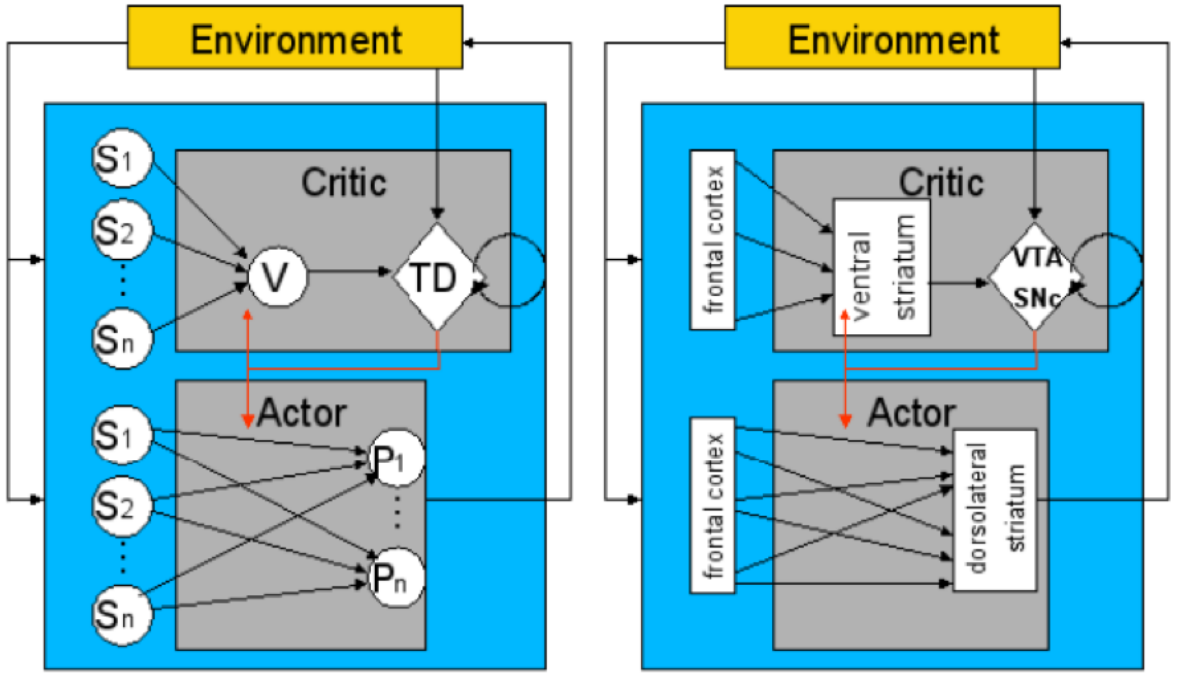
Schultz et al. (1993);
Houk et al. (1995); Schultz et al. (1997).

$$\overbrace{\delta_{t+1}}^{+1} = r_{t+1} + \gamma \cdot V(s_{t+1}) - V(s_t)$$

$$\overbrace{\delta_{t+1}}^{0} = r_{t+1} + \gamma \cdot V(s_{t+1}) - V(s_t)$$

$$\overbrace{\delta_{t+1}}^{-1} = r_{t+1} + \gamma \cdot V(s_{t+1}) - V(s_t)$$

Actor-Critic model of the basal ganglia



From Takahashi, Schoenbaum and Niv, Frontiers in Neurosciences, pp. 86-97, july 2008

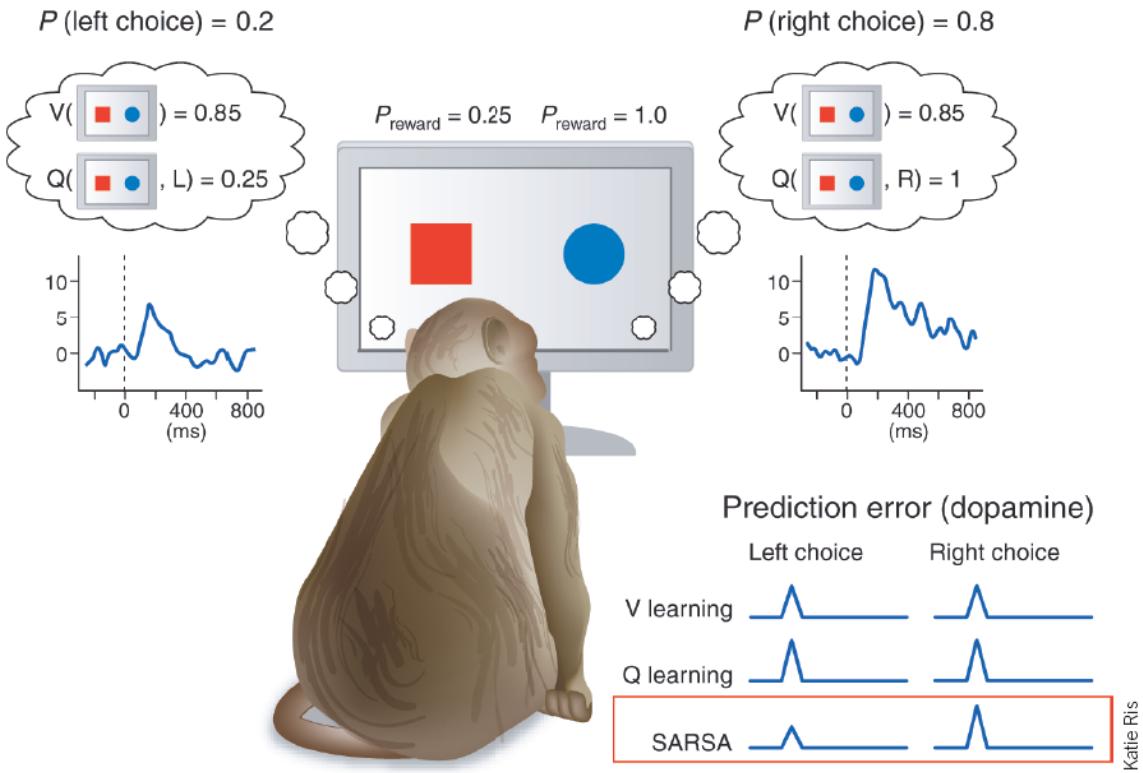
The Actor learns to select actions that maximize reward.
The Critic learns to predict reward (its value V).
A **reward prediction error** signal constitutes the reinforcement signal.
Also see [Khamassi et al. 2005].

Which TD-learning algorithm is consistent with dopamine activity?

- V-learning (e.g., Actor-Critic):
  - $V(s_{t-1}) = V(s_{t-1}) + \alpha[r_t + \gamma V(s_t) - V(s_{t-1})]$
  - $P(a_{t-1}|s_{t-1}) = P(a_{t-1}|s_{t-1}) + \alpha_A[r_t + \gamma V(s_t) - V(s_{t-1})]$
- Q-learning:
  - $Q(s_{t-1}, a_{t-1}) = Q(s_{t-1}, a_{t-1}) + \alpha[r_t + \gamma \max_a Q(s_t, a) - Q(s_{t-1}, a_{t-1})]$
- SARSA:
  - $Q(s_{t-1}, a_{t-1}) = Q(s_{t-1}, a_{t-1}) + \alpha[r_t + \gamma Q(s_t, a_t) - Q(s_{t-1}, a_{t-1})]$

[Sutton & Barto 1998]

Which TD-learning algorithm is consistent with dopamine activity when animals make a choice?



Niv et al. (2006), commentary about the results presented in Morris et al. (2006).

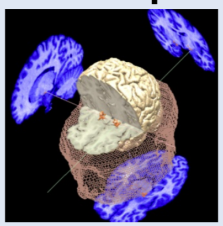
Popularity of RL models in Neuroscience of decision-making

Sequence of observed trials : Left (Reward); Left (Nothing); Right (Nothing); Left (Reward); …

fMRI scanner

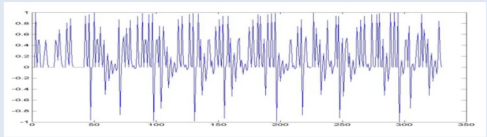
RL model

Brain responses



?

Reward prediction error (δ)



e.g. Palminteri Khamassi Joffily Coricelli (2015) Nature Communications

Adapted from a slide by Peter Dayan

Limitations

This is a nice story, but there are some limitations

- Most tasks are single-step (but see Daw et al., 2011)
- Very small number of discrete states and actions
- Assumed perfect state identification
- Animals sometimes learn fast
- In parallel, applications to Robotics gave disappointing results.

Animal fast adaptation in some situations

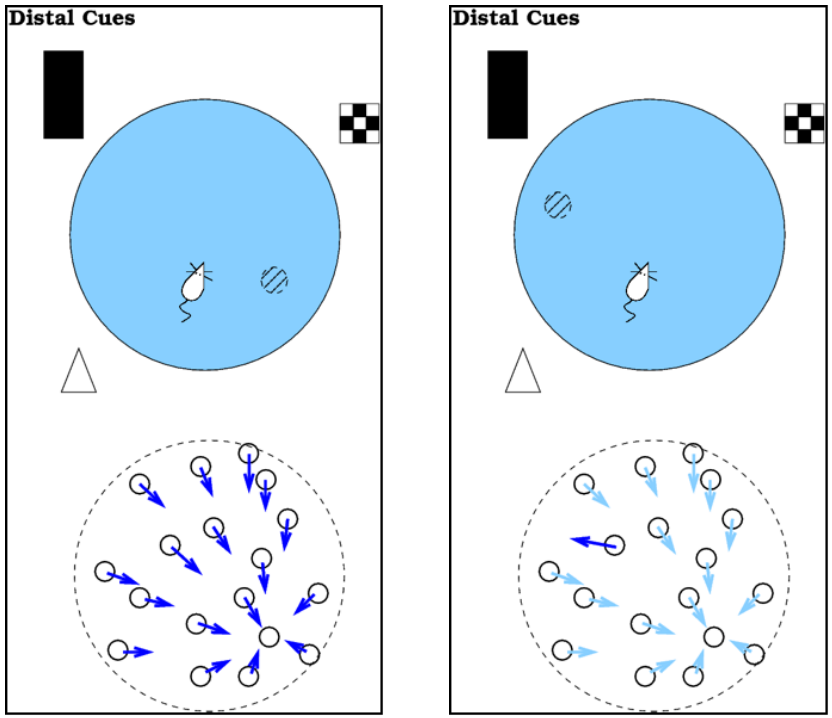
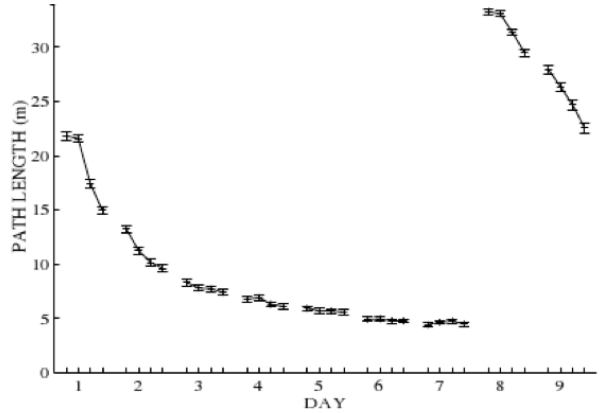
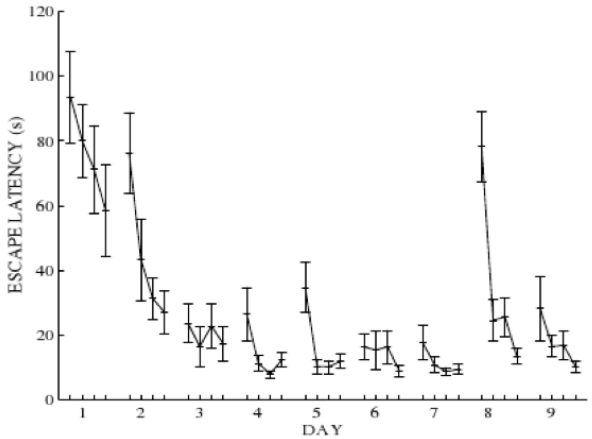


Figure by Benoît Girard (ISIR / Sorbonne).



Simulation of MF-RL by Foster et al. (2000)



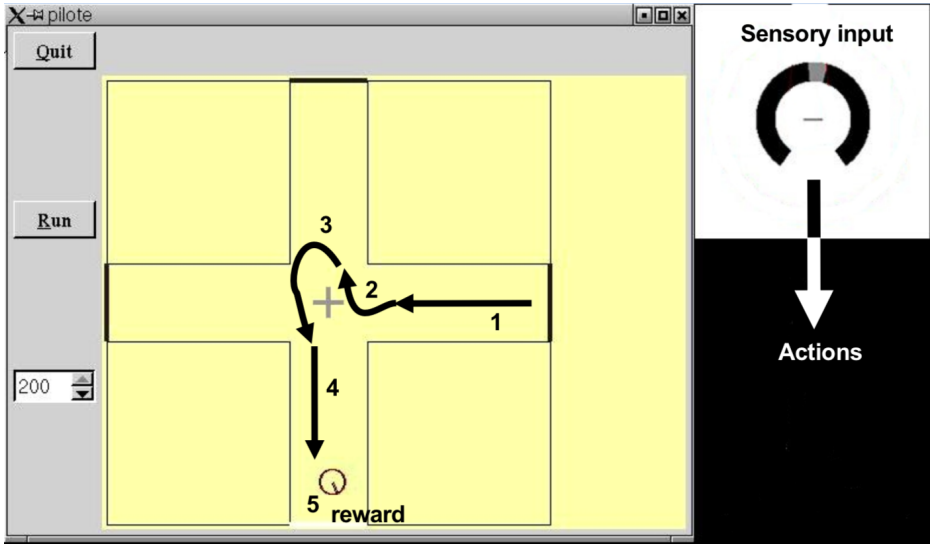
Animal behavior in Morris (1982)

Applications of MF-RL to Robotics

- **Smart & Kaelbling 2002**: Requires initial trajectory demonstration by the human.
- **Morimoto & Doya 2001**: Efficient but unstable.
- **Sporns & Alexander 2002**: Simple discrete task.
- **Arleo et al. 2004; Krichmar et al. 2005; Khamassi et al. 2006**: Requires an important step for state decomposition.
- **ALL**: Slow learning. Local optima. Prior knowledge.

- **BUT**: See work by Peters & Schaal 2006, 2008 to learn model-free continuous motor primitives. Also the parameterized RL framework combining continuous and discrete action spaces [Khamassi et al. 2018 IEEE Trans Cog Dev Sys].

Continuous reinforcement learning (I)



MF-RL applied to navigation behavior learning in a robot performing the bio-inspired plus-maze task [Khamassi et al., 2005 Adaptive Behavior (Journal)].

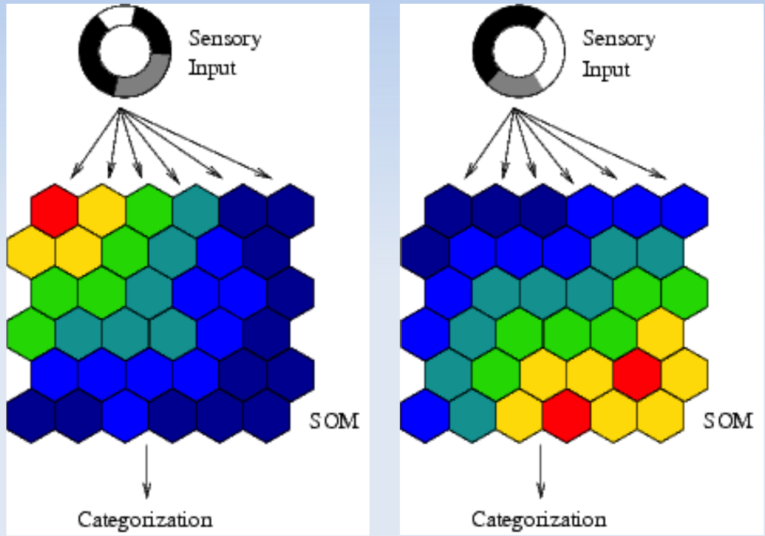
Continuous reinforcement learning (II)

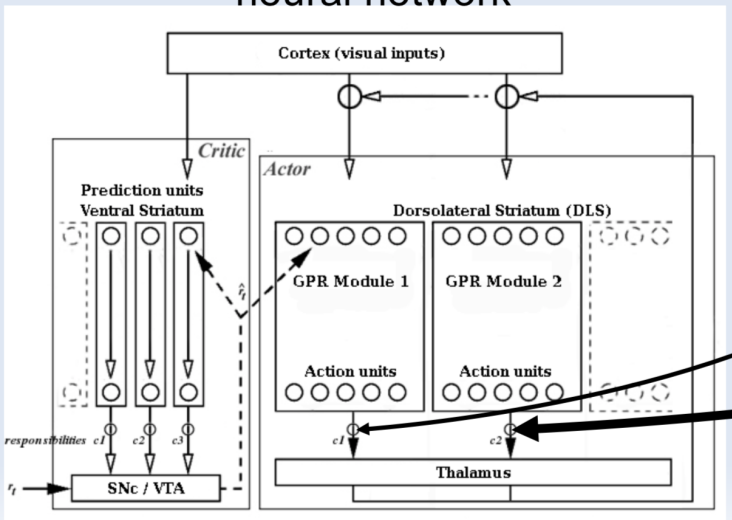
Unsupervised Learning → Coordination by a self-organizing map

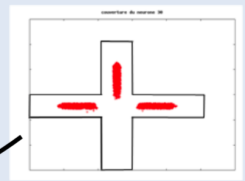
Reinforcement Learning

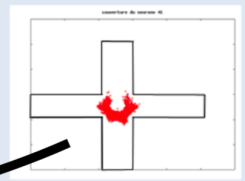
↓

Actor-Critic multi-modules neural network



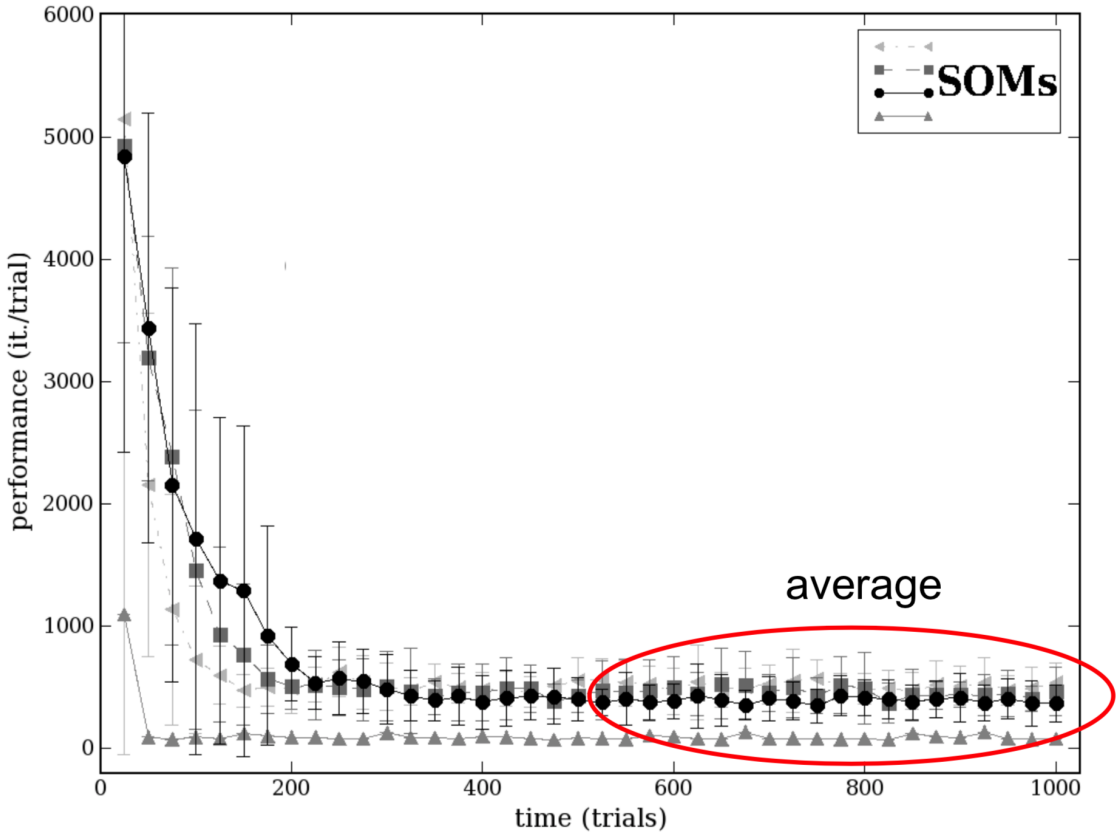






Khamassi et al., 2006 International Conference on Adaptive Behavior

Continuous reinforcement learning (III)



Khamassi et al., 2006 International Conference on Adaptive Behavior

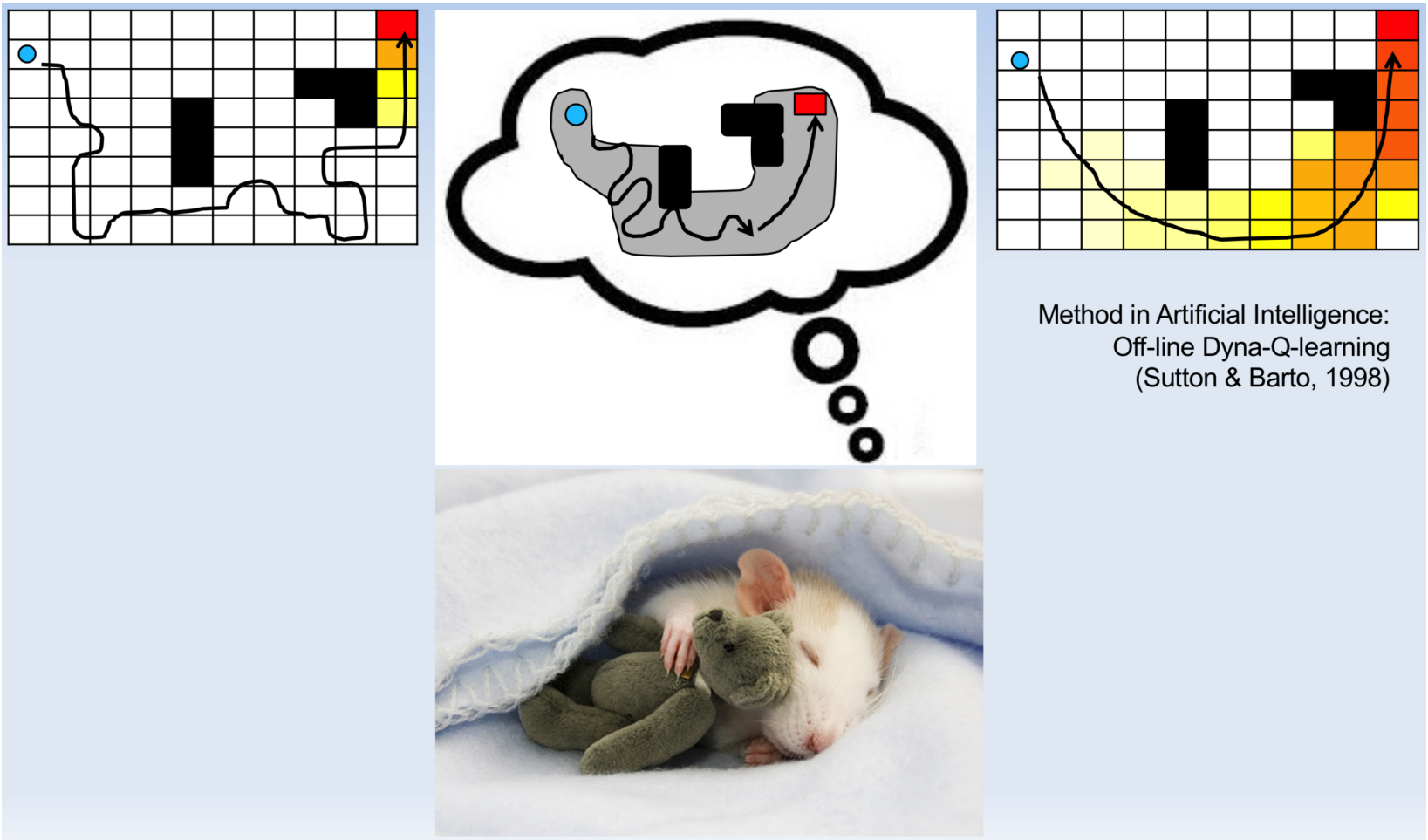
Model-based reinforcement learning

Model-based reinforcement learning

- A **model-based (MB) agent** learns an estimate of the two functions that define a *model* of the task:
  - The reward function, $\hat{R} : (S, A) \rightarrow \mathbb{R}$.
  - The transition function, $\hat{T} : (S, A) \rightarrow \Pi(S)$.
- A classical way to learn the model consists in measuring the frequency of state and reward observations following each encountered (state,action) couple.
- A classical way to learn the (state,action) value function from the model is **dynamic programming/value iteration**:
  - $Q(s, a) = \hat{R}(s, a) + \gamma \sum_{s'} \hat{T}(s'|s, a) max_{a'} Q(s', a')$

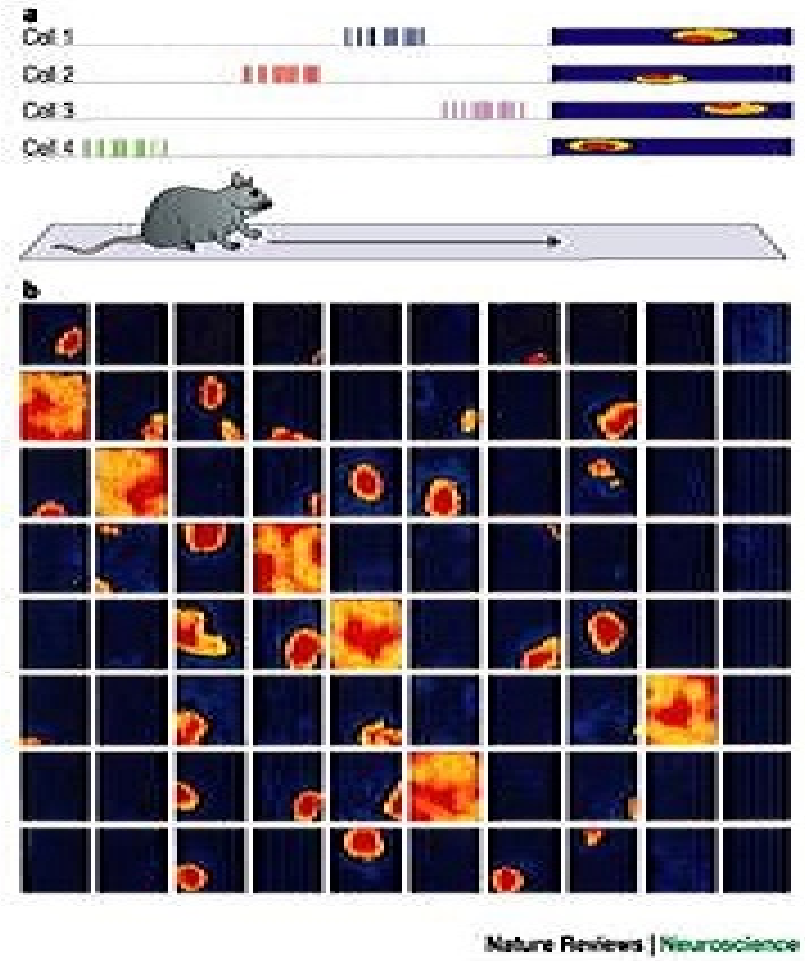
[Sutton & Barto 1998]

Model-based reinforcement learning during “sleep”



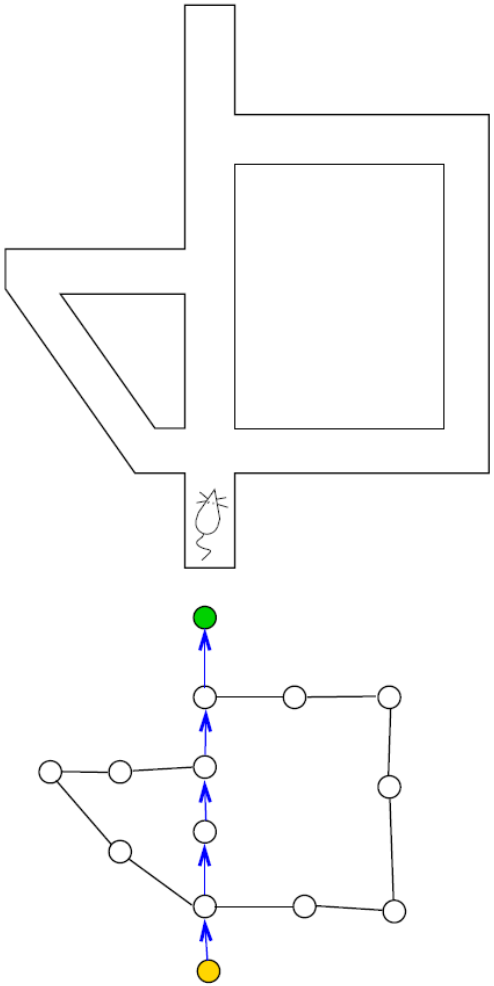
Cazé*, Khamassi* et al., (2018) Journal of Neurophysiology

Hippocampal place cells



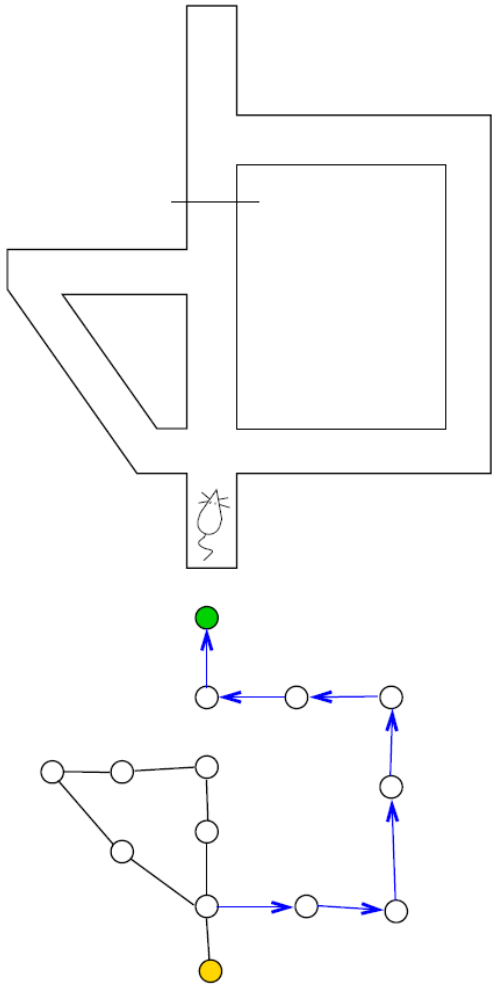
Nakazawa, McHugh, Wilson, Tonegawa (2004) Nature Reviews Neuroscience

Model-based decision-making (planning) in rats



Martinet et al. (2011) model applied to the Tolman maze

Model-based decision-making (planning) in rats



Martinet et al. (2011) model applied to the Tolman maze

Hippocampal activity during deliberation in rats

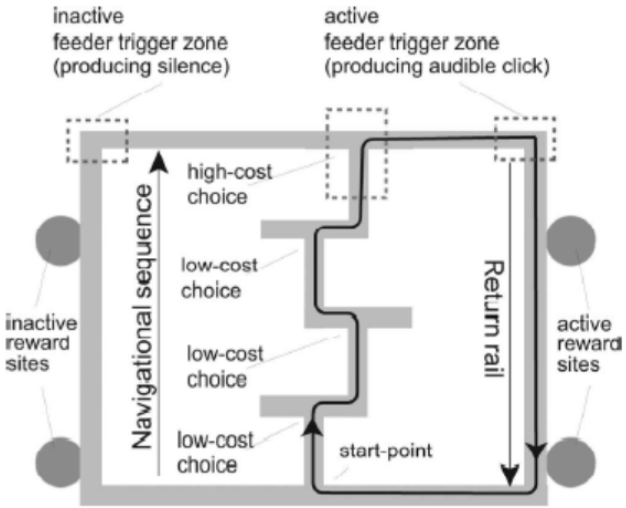
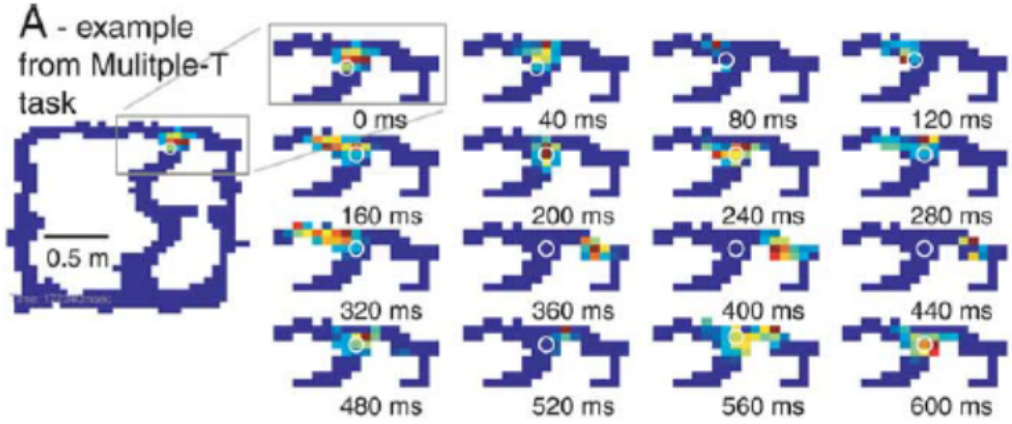
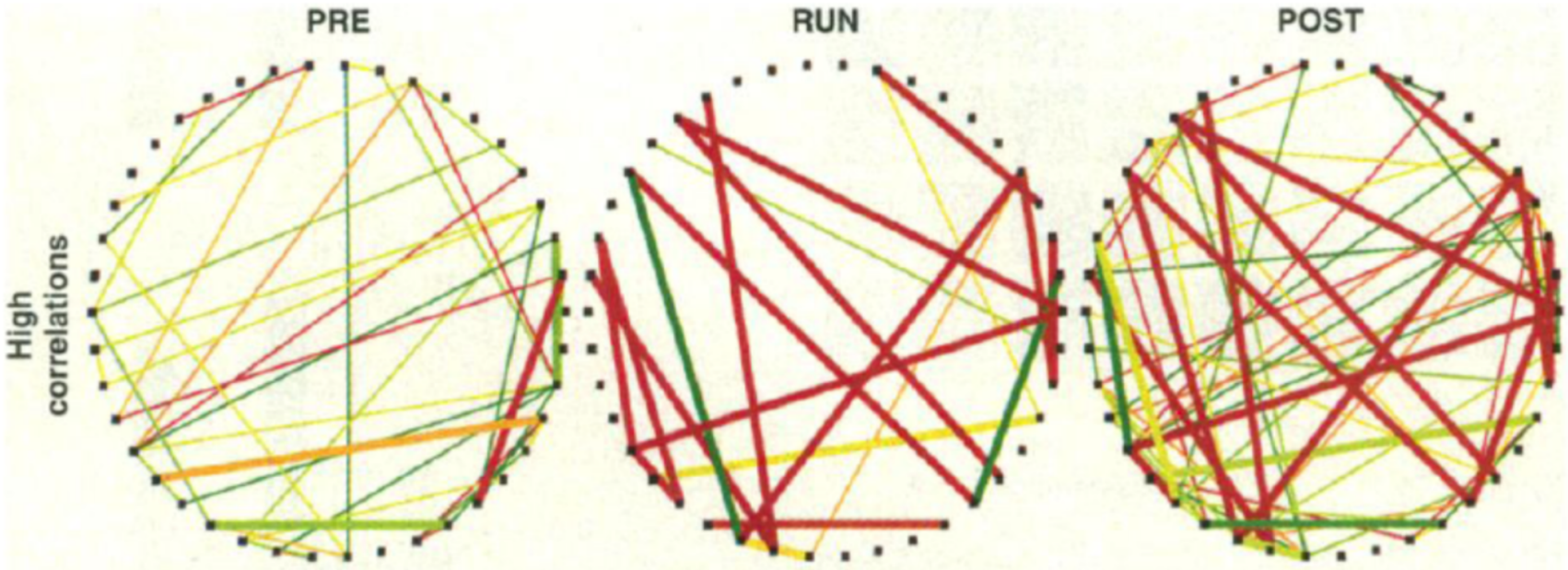


Figure 1. The multiple-T maze. The task consists of four T choice points with food reward available at two sites on each return rail. Only feeders on one side of the track were rewarded in each session.



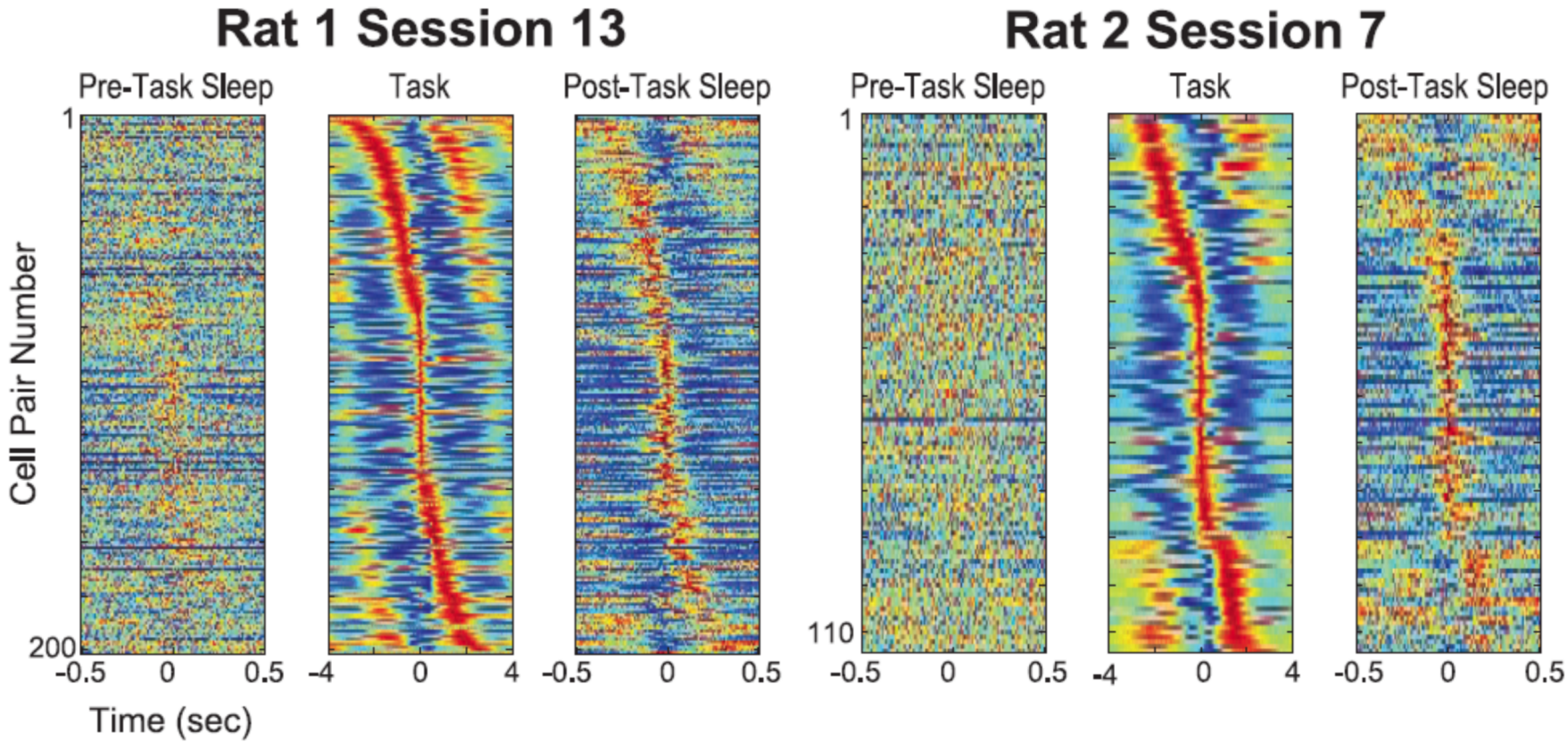
Johnson & Redish (2007) Journal of Neuroscience

Hippocampal place cells



Reactivation of hippocampal place cells during sleep (Wilson & McNaughton, 1994)

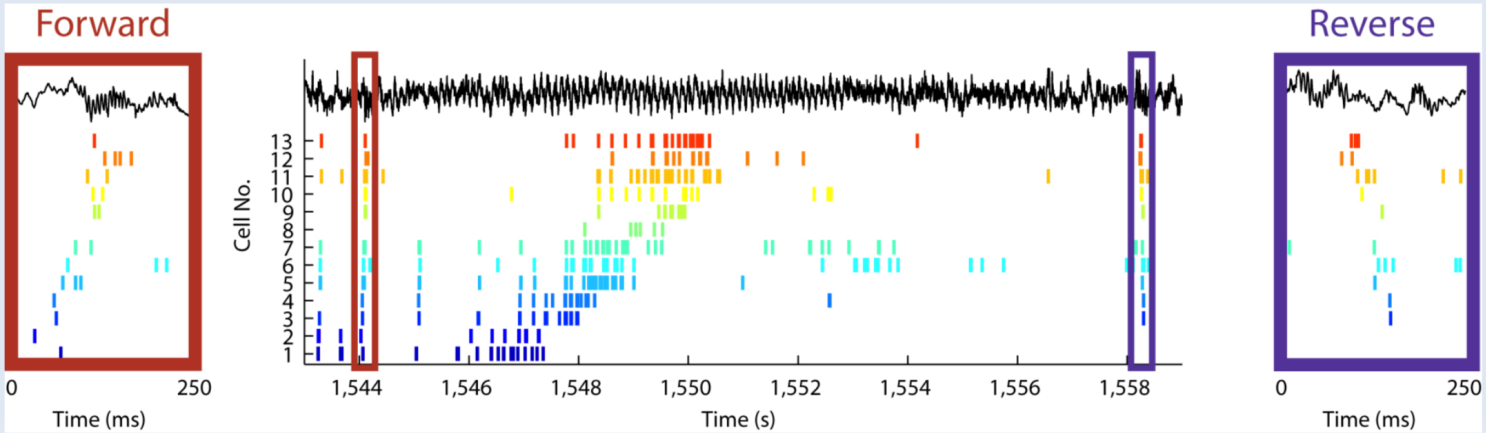
Replay also in the Prefrontal cortex



Forward replay of prefrontal cortex neurons during sleep (sequence is compressed 7 times) (Euston et al., 2007, Science)

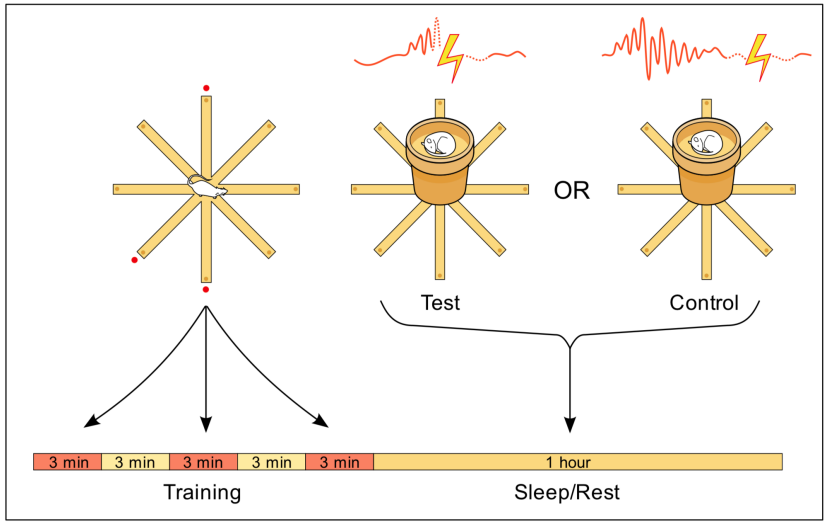
Hippocampal place cells

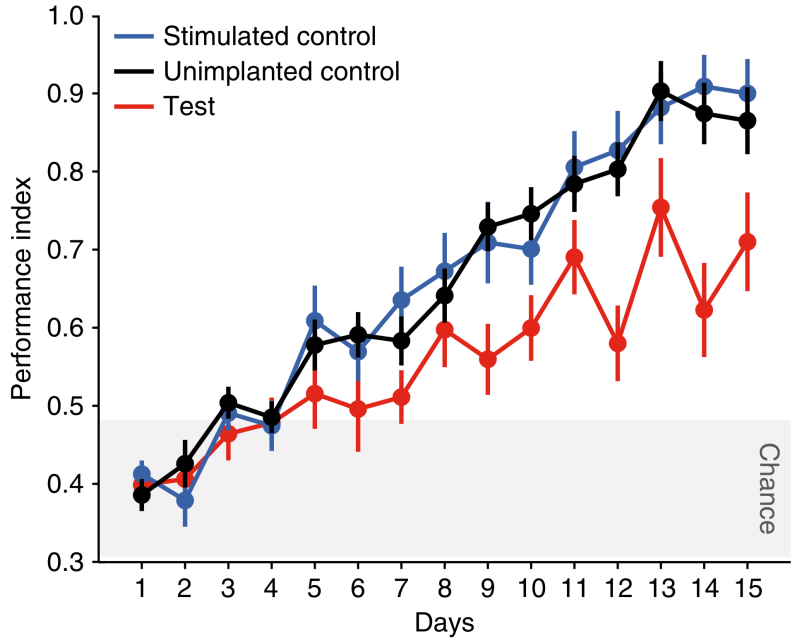
"Ripple" events = irregular bursts of population activity that give rise to brief but intense high-frequency (100-250 Hz) oscillations in the CA1 pyramidal cell layer.



Diba & Buszaki (2007)

Causal role for SWRs in learning



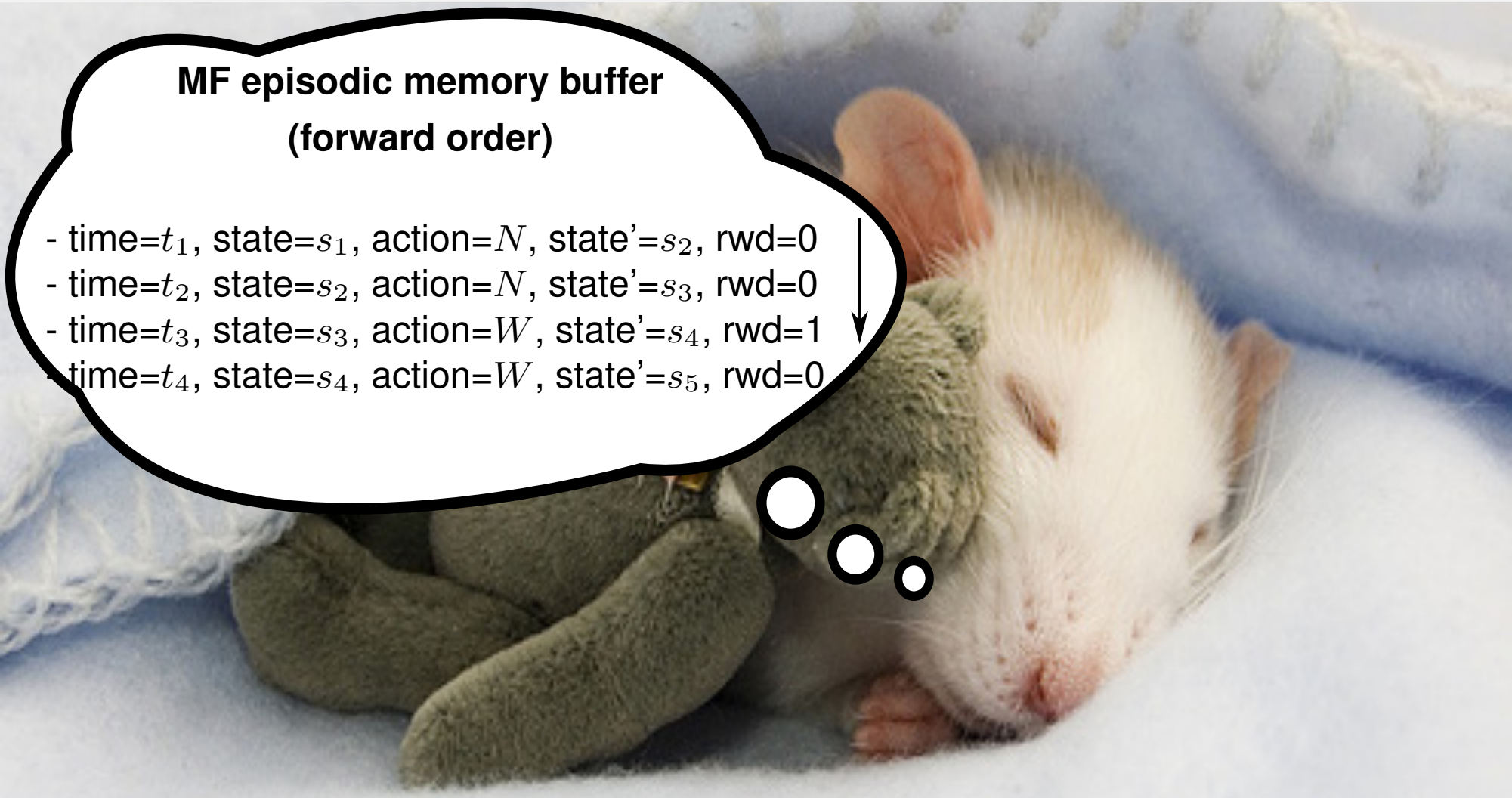


Girardeau G, Benchenane K, Wiener SI, Buzsáki G, Zugaro MB (2009)

Reactivation (replay) (MF) vs. mental simulation (MB)

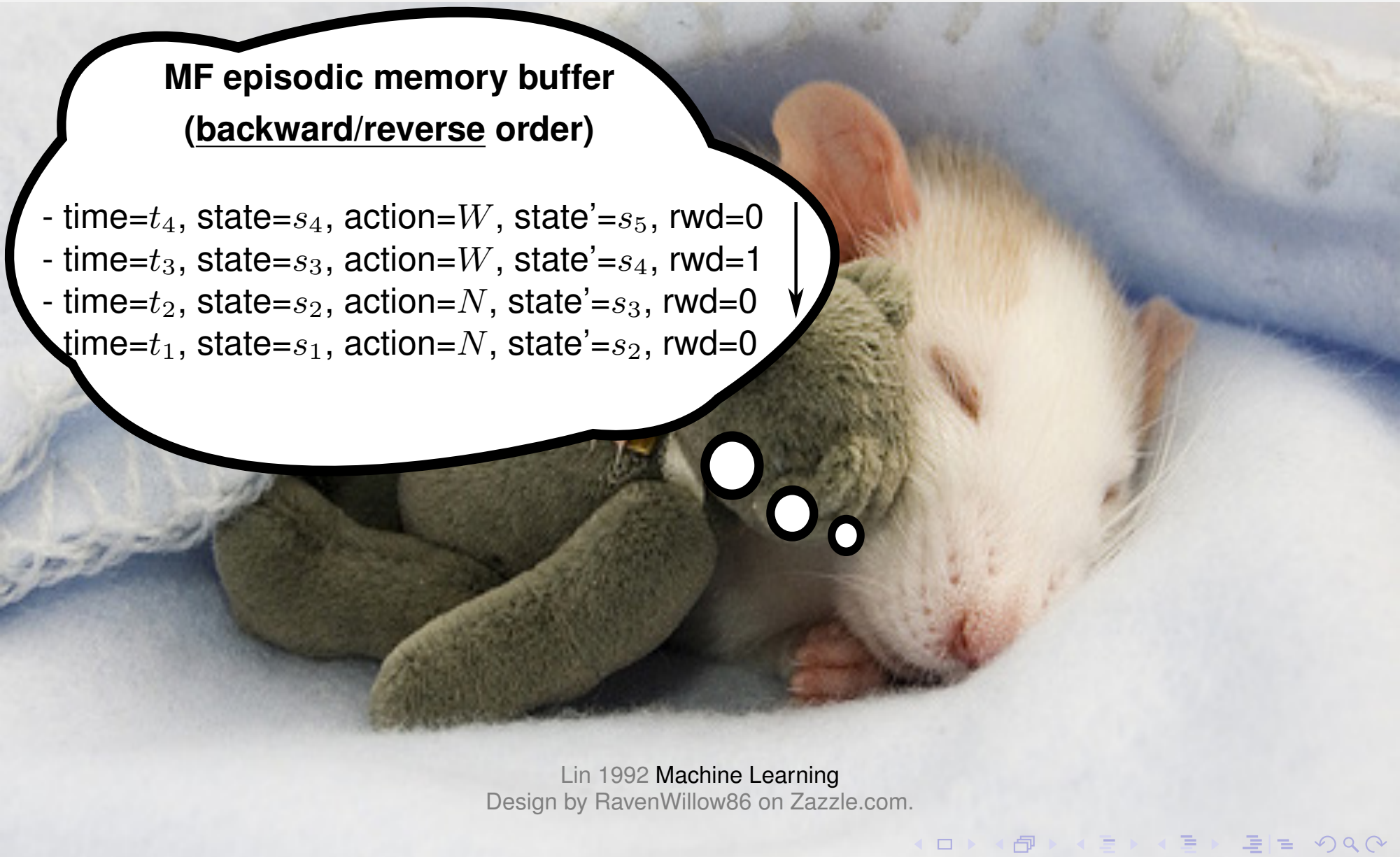
Design by RavenWillow86 on Zazzle.com.

Reactivation (replay) (MF) vs. mental simulation (MB)



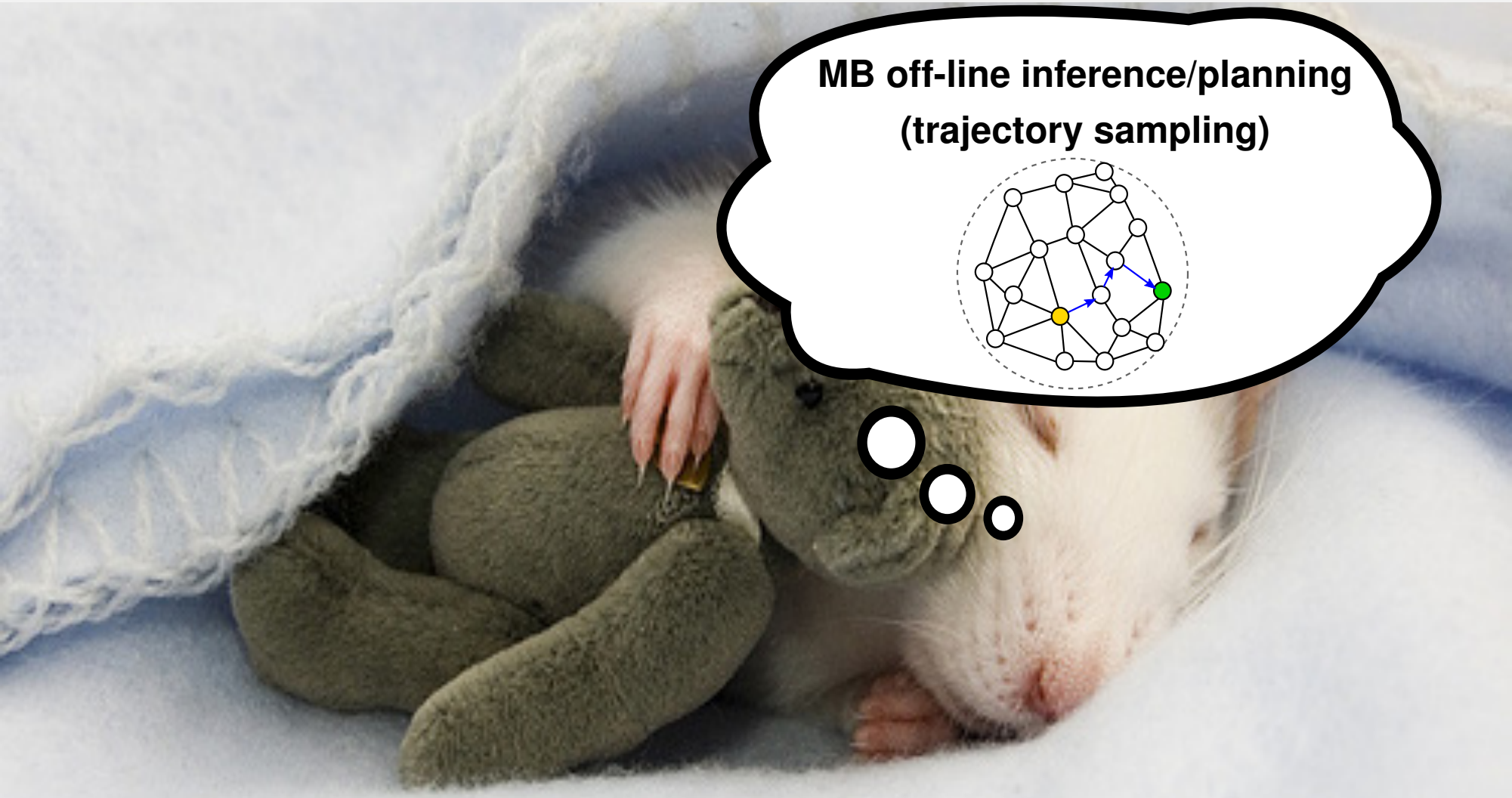
Caze* Khamassi* Aubin Girard 2018 Journal of Neurophysiology
Design by RavenWillow86 on Zazzle.com.

Reactivation (replay) (MF) vs. mental simulation (MB)



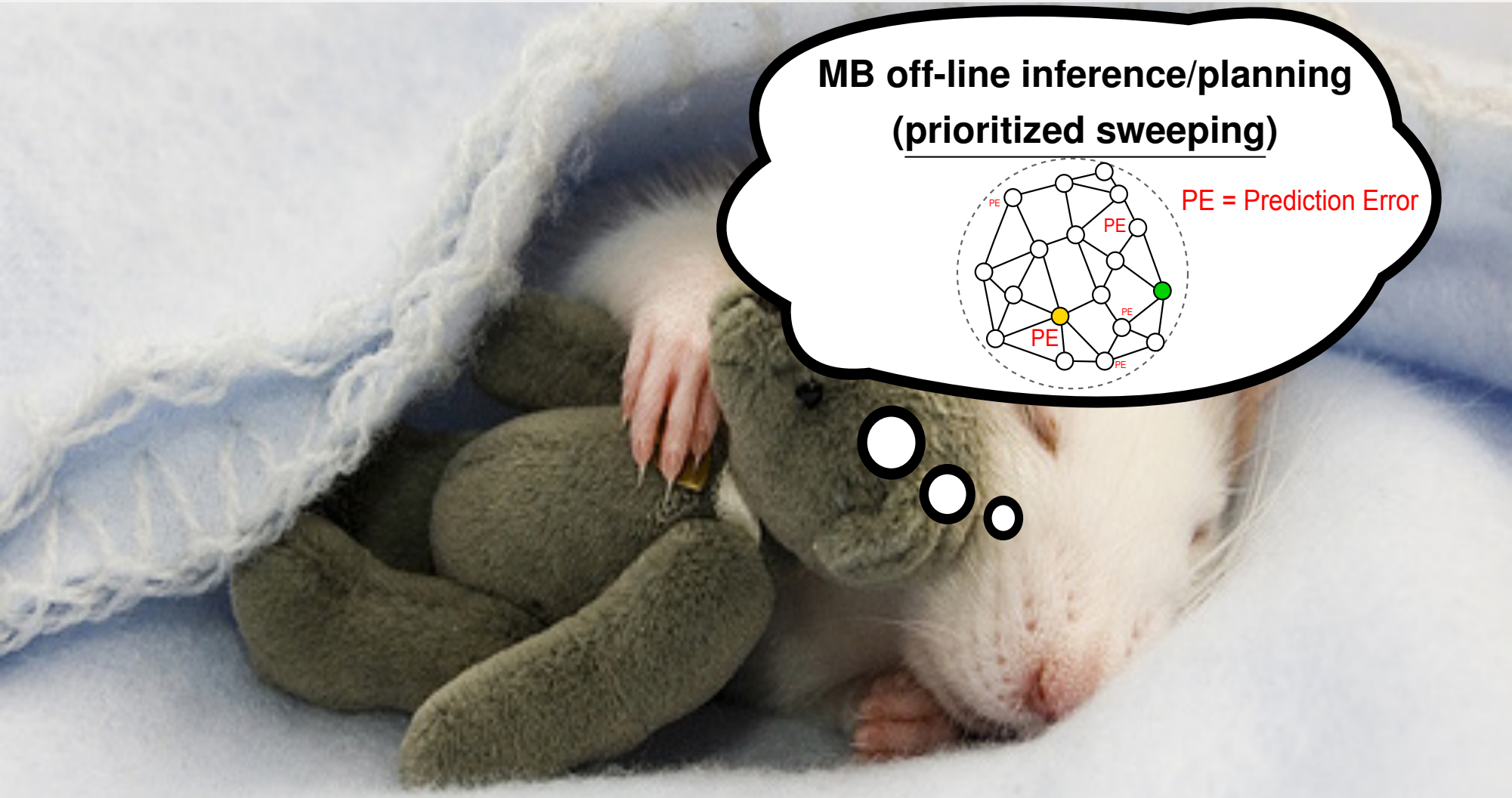
Lin 1992 Machine Learning
Design by RavenWillow86 on Zazzle.com.

Reactivation (replay) (MF) vs. mental simulation (MB)



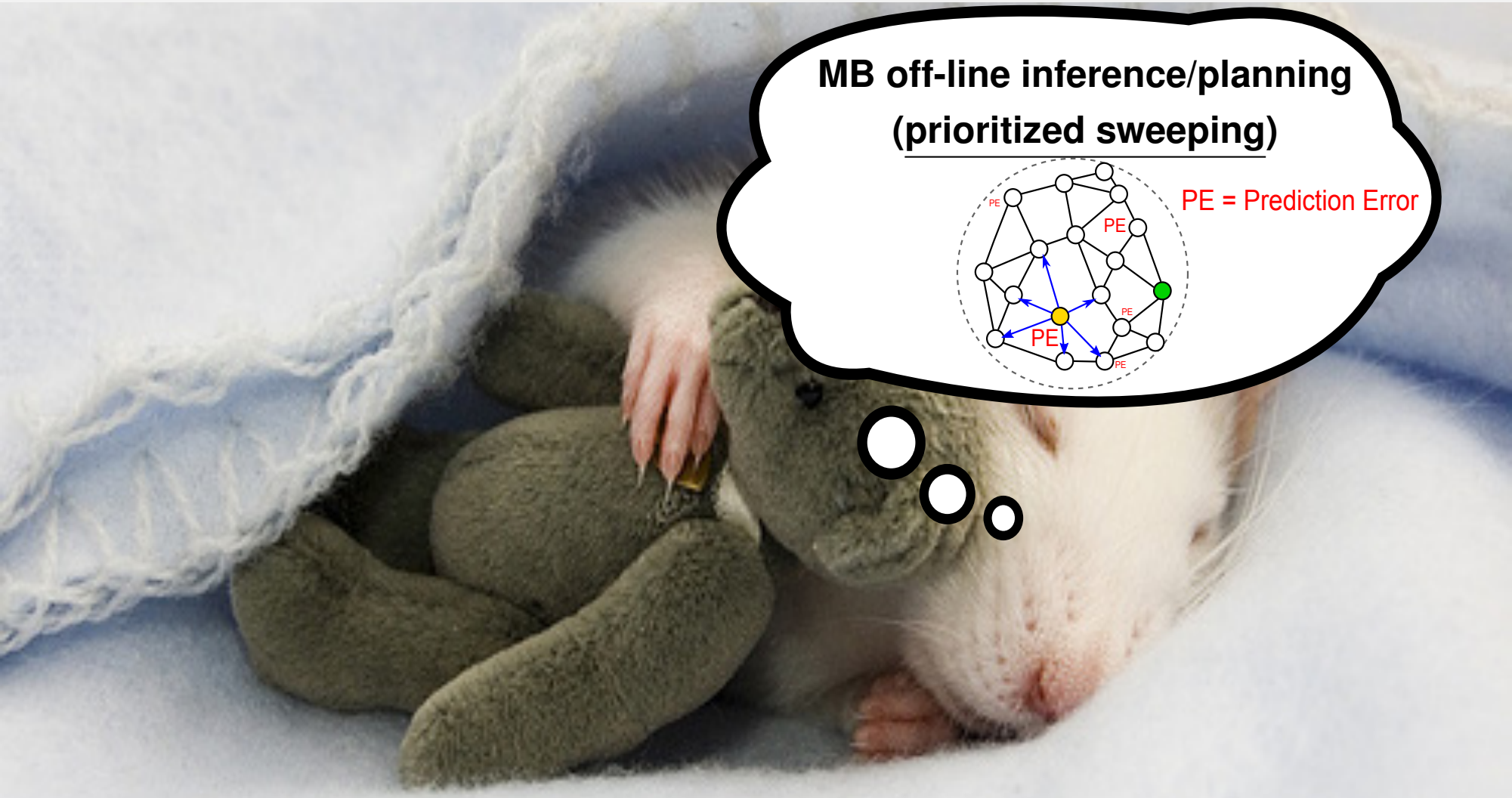
Khamassi Girard 2020 Biological Cybernetics
Design by RavenWillow86 on Zazzle.com.

Reactivation (replay) (MF) vs. mental simulation (MB)



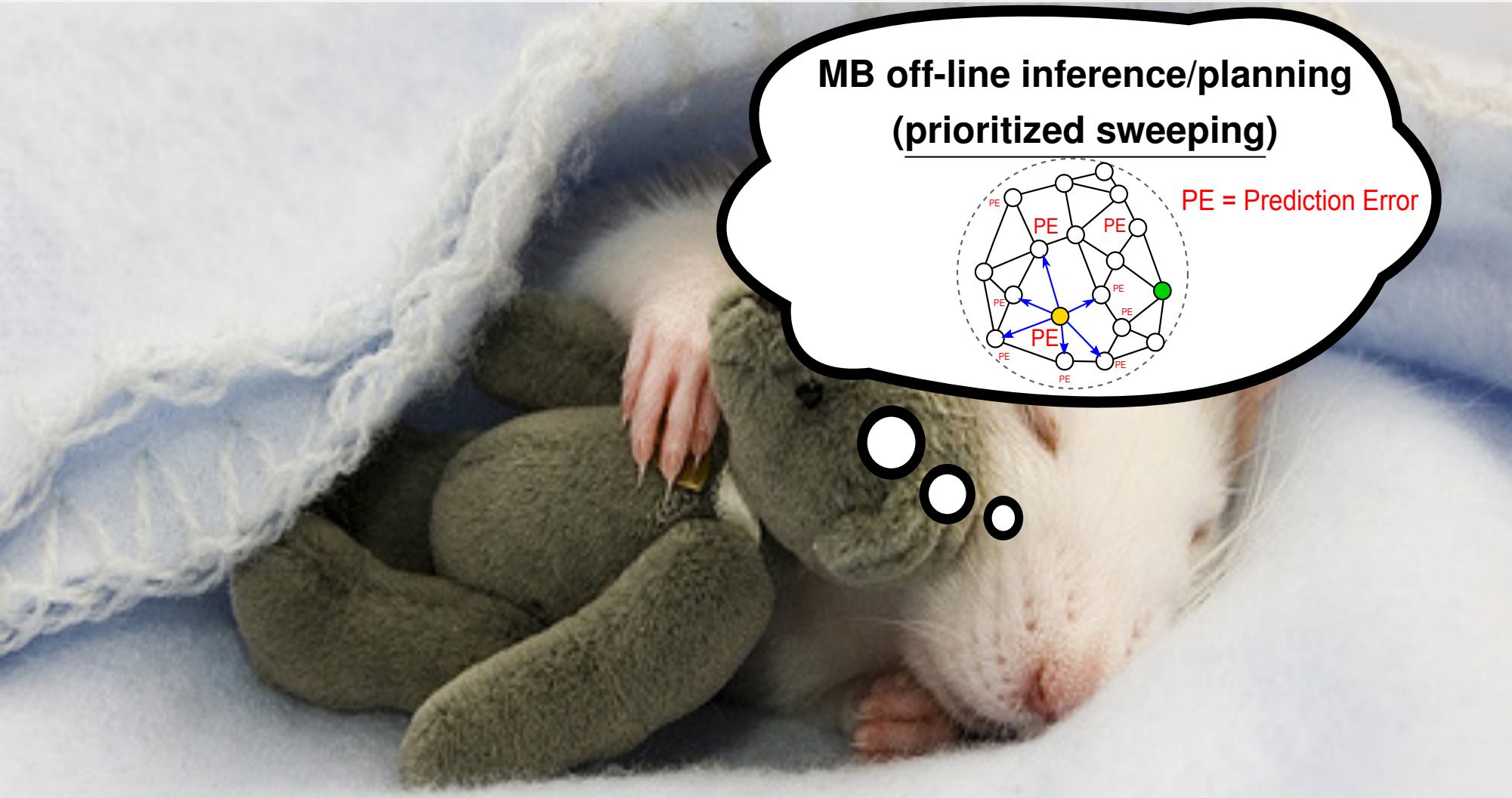
Khamassi Girard 2020 Biological Cybernetics
Design by RavenWillow86 on Zazzle.com.

Reactivation (replay) (MF) vs. mental simulation (MB)



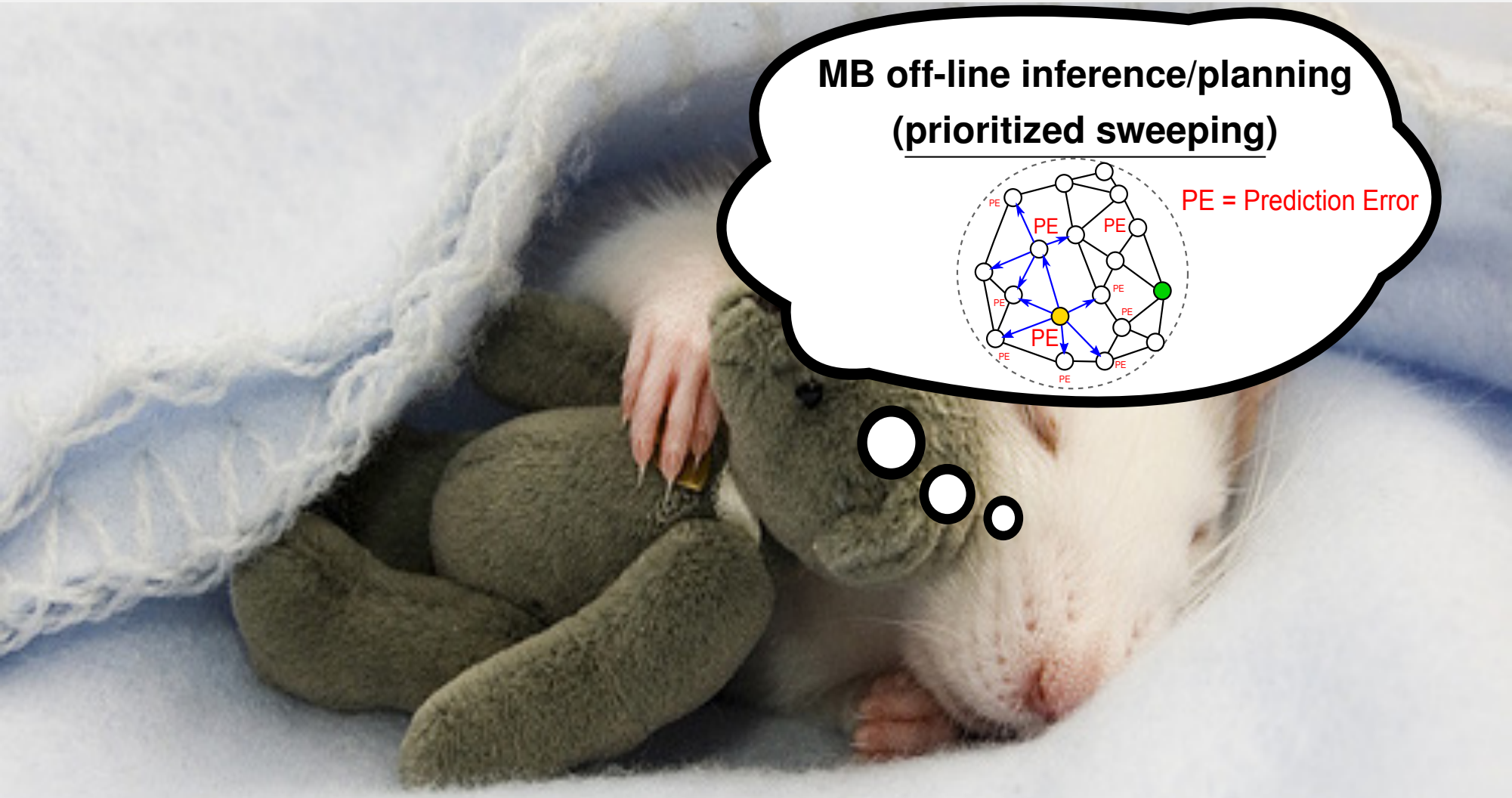
Khamassi Girard 2020 Biological Cybernetics
Design by RavenWillow86 on Zazzle.com.

Reactivation (replay) (MF) vs. mental simulation (MB)



Khamassi Girard 2020 Biological Cybernetics
Design by RavenWillow86 on Zazzle.com.

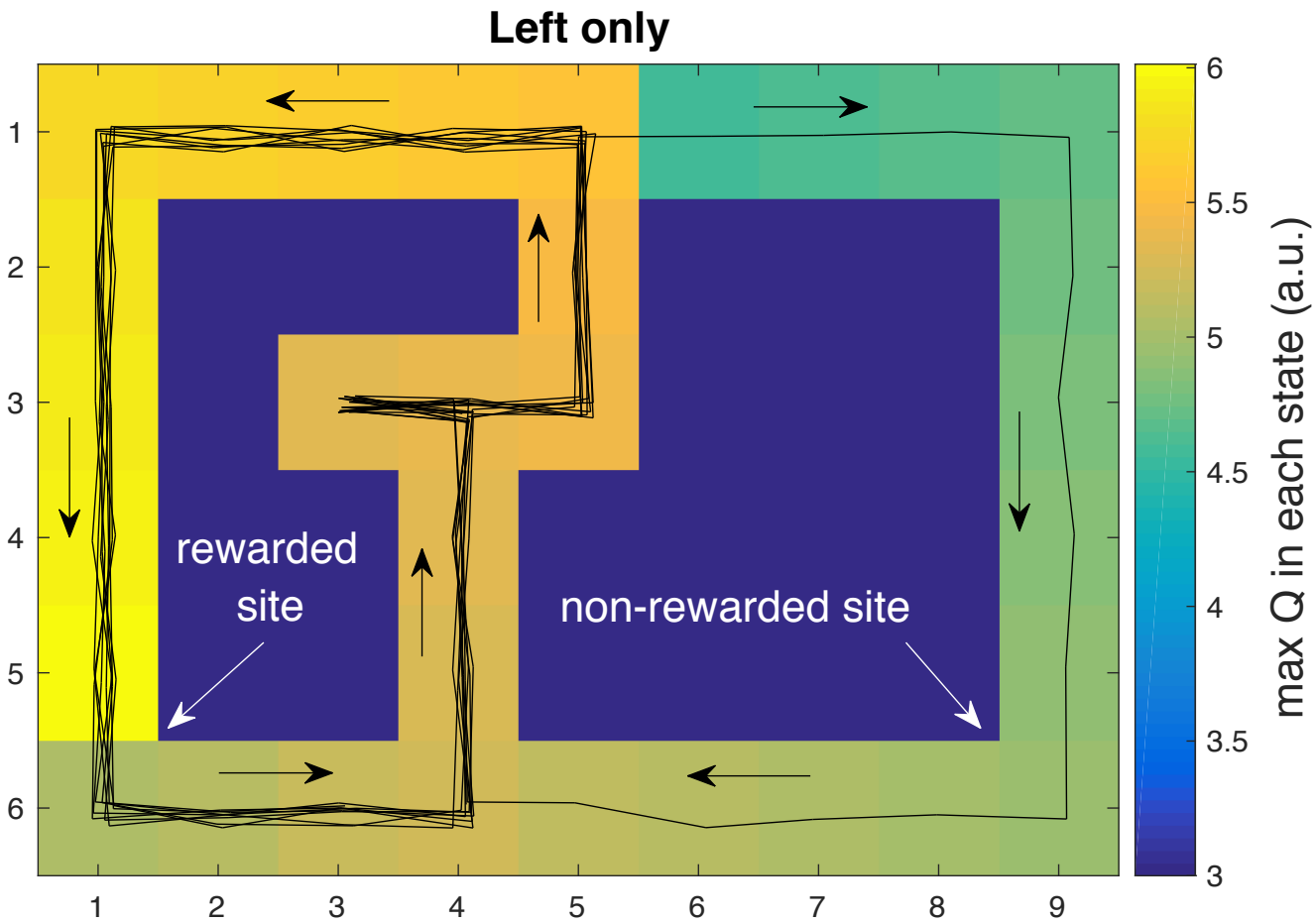
Reactivation (replay) (MF) vs. mental simulation (MB)



Khamassi Girard 2020 Biological Cybernetics
Design by RavenWillow86 on Zazzle.com.

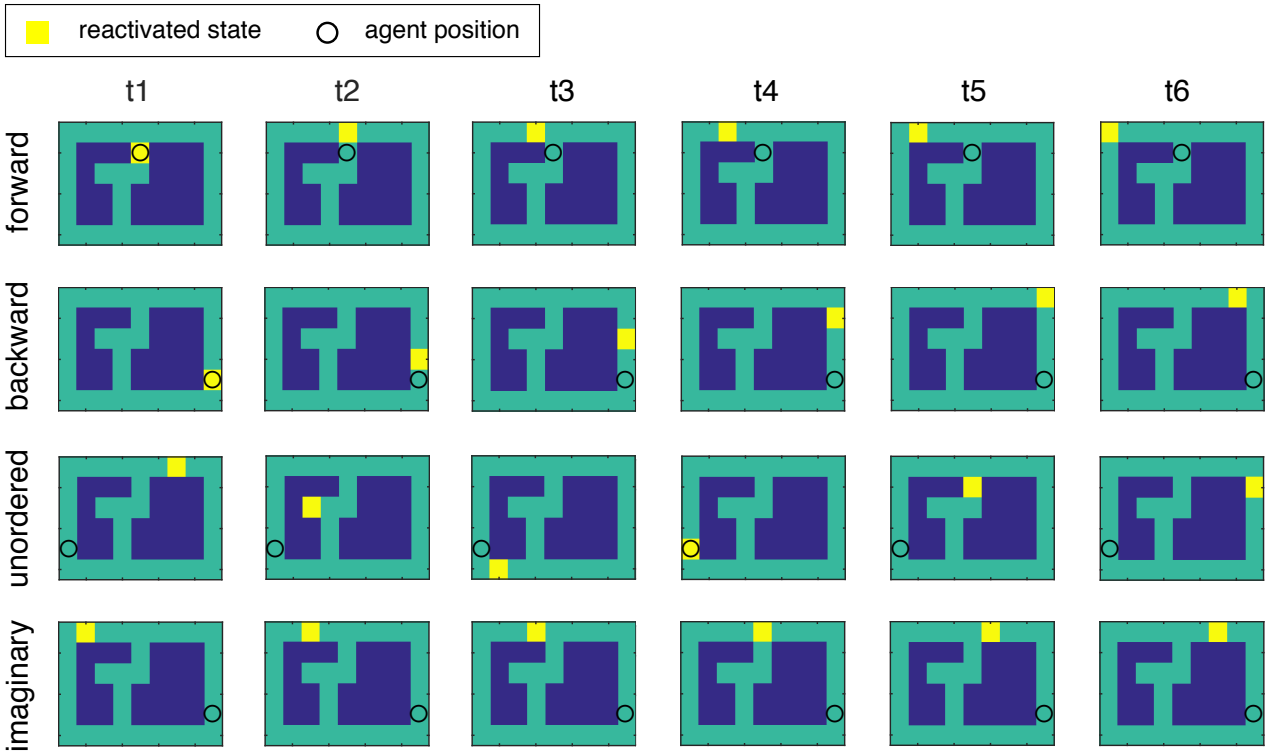
Replay in MB/MF reinforcement learning

Example of a discrete grid-world navigation task



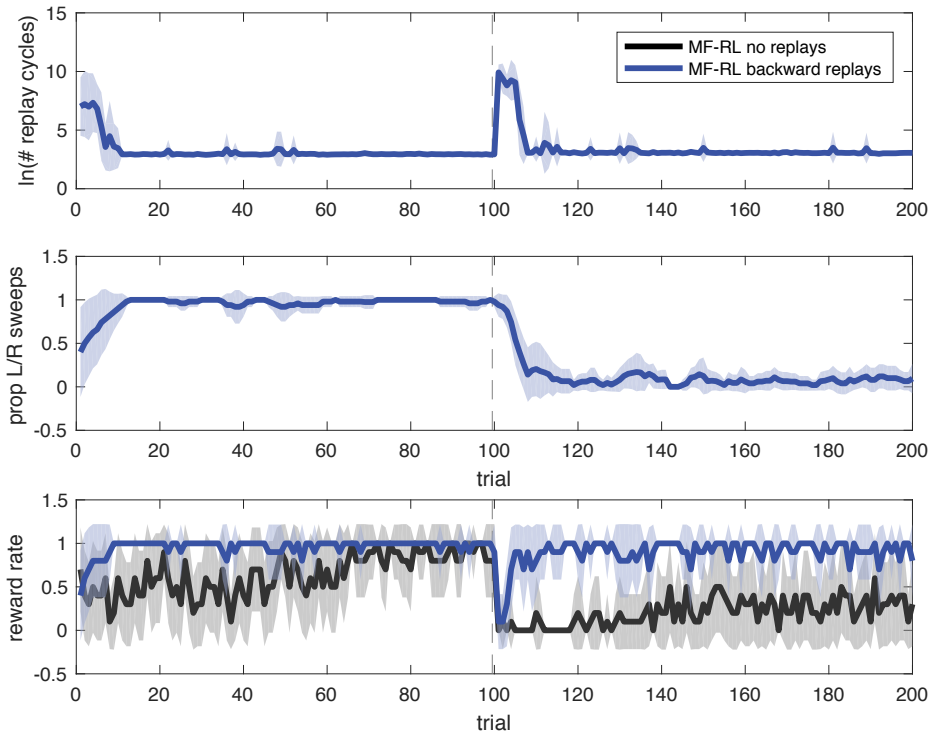
[Caze*, Khamassi* et al. 2018 J Neurophysiol]
Q-values learned by a model-free RL agent (here with backward replay).

MB/MF RL off-line replay/reactivations



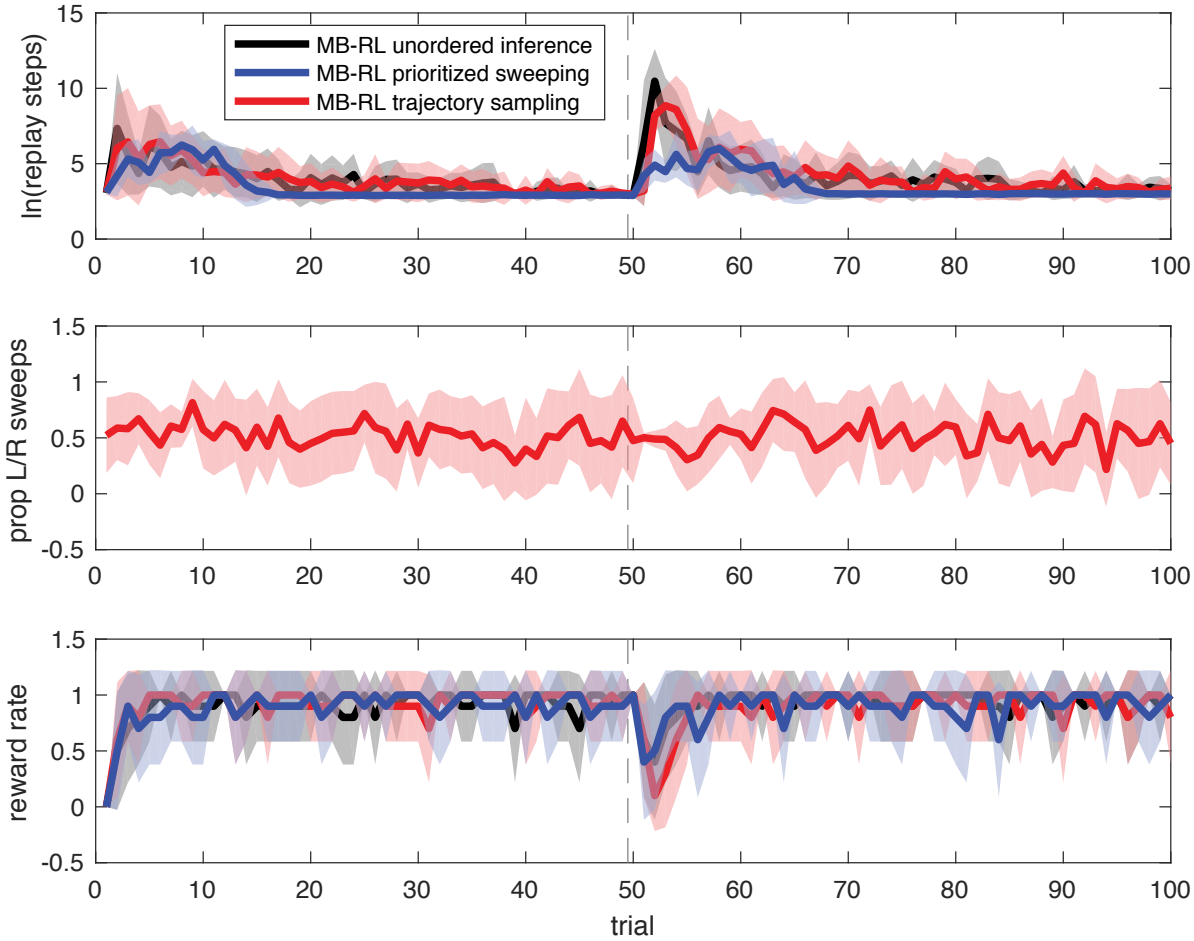
Cazé*,Khamassi* et al. (2018) J Neurophysiology. Khamassi & Girard (2020)
https://github.com/MehdiKhamassi/RLwithReplay

MB/MF RL off-line replay/reactivations



Cazé*,Khamassi* et al. (2018) J Neurophysiology. Khamassi & Girard (2020)
https://github.com/MehdiKhamassi/RLwithReplay

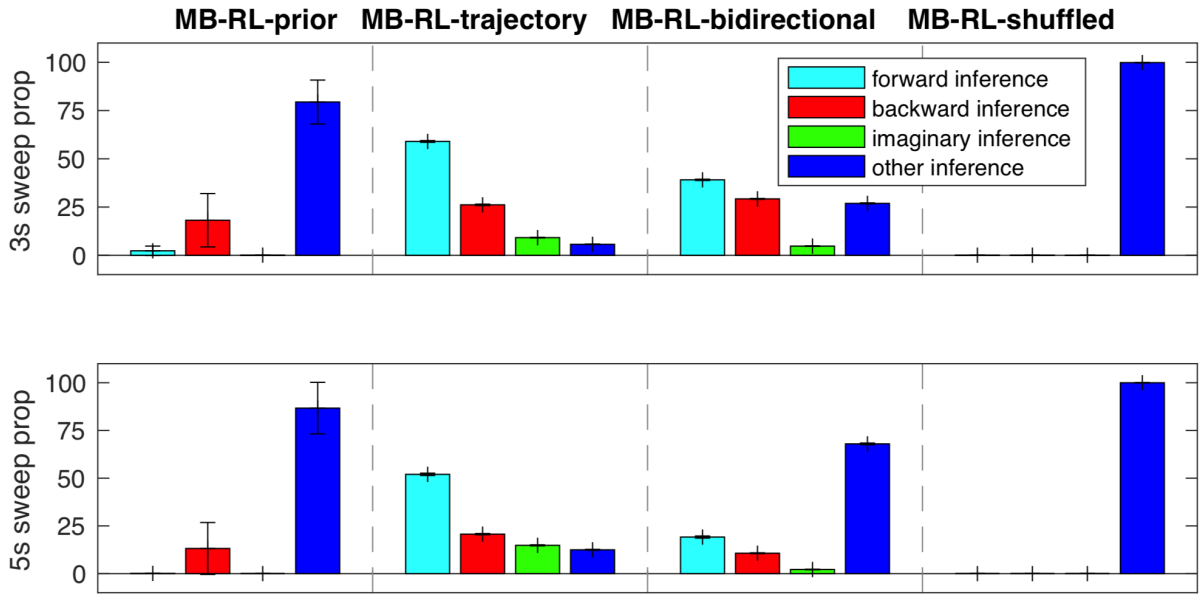
MB/MF RL off-line replay/reactivations



Cazé*,Khamassi* et al. (2018) J Neurophysiology. Khamassi & Girard (2020)
https://github.com/MehdiKhamassi/RLwithReplay

MB/MF RL off-line replay/reactivations

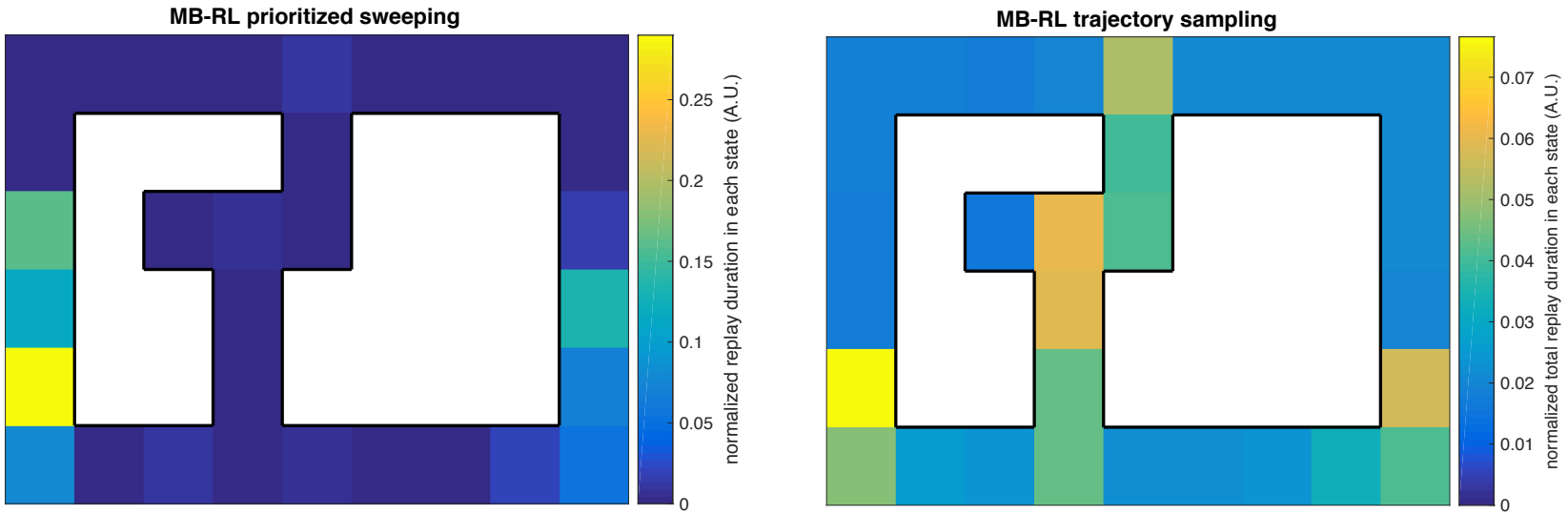
Different models predict different proportions of forward/backward/random replay



Cazé*,Khamassi* et al. (2018) J Neurophysiology. Khamassi & Girard (2020)
https://github.com/MehdiKhamassi/RLwithReplay

MB/MF RL off-line replay/reactivations

Different models predict different locations where to stop to perform replay



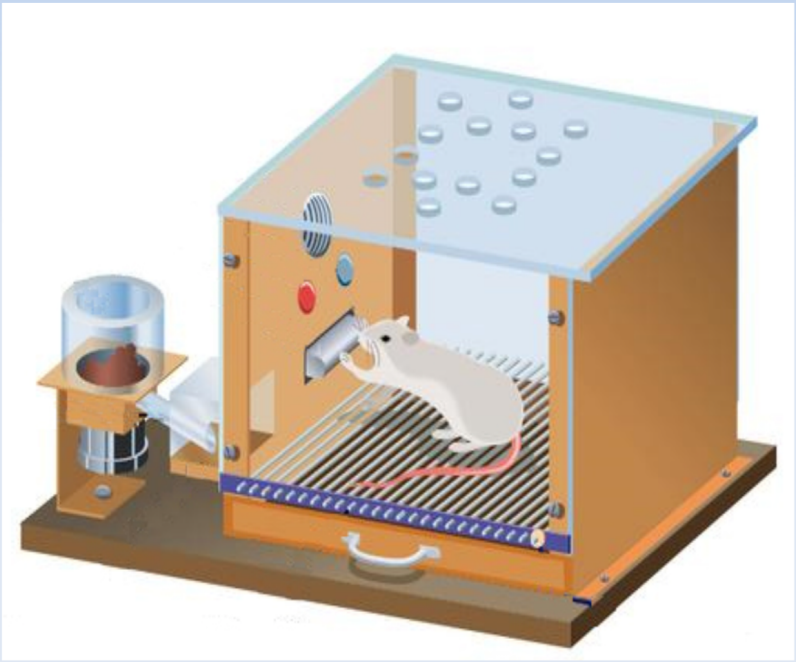
Replay at reward site vs. replay at decision-point

Caze* Khamassi* Aubin Girard 2018 Journal of Neurophysiology

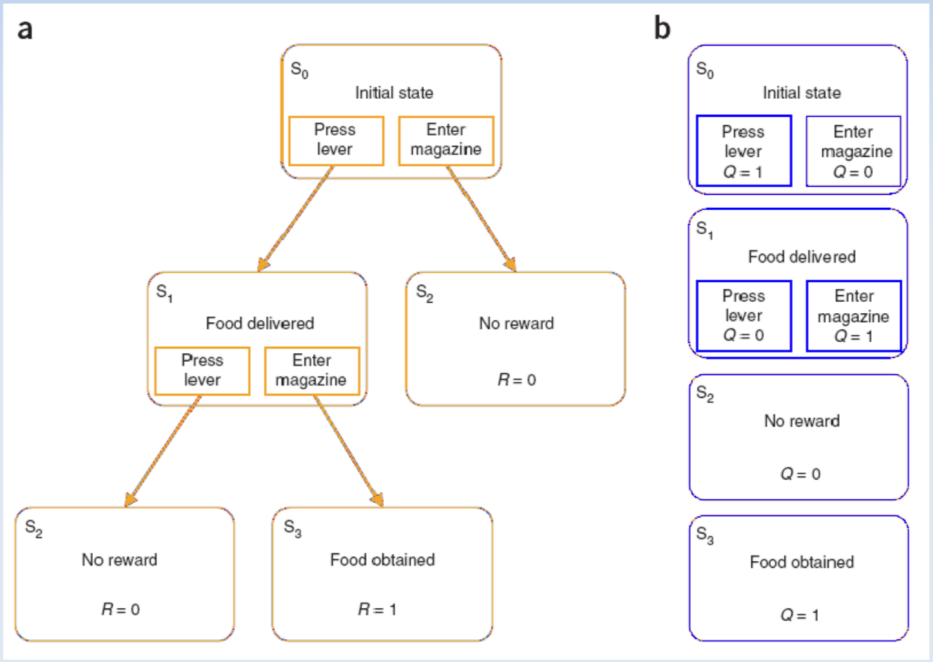
MB/MF combination

Multiple decision systems in rats

Skinner box (instrumental conditioning)



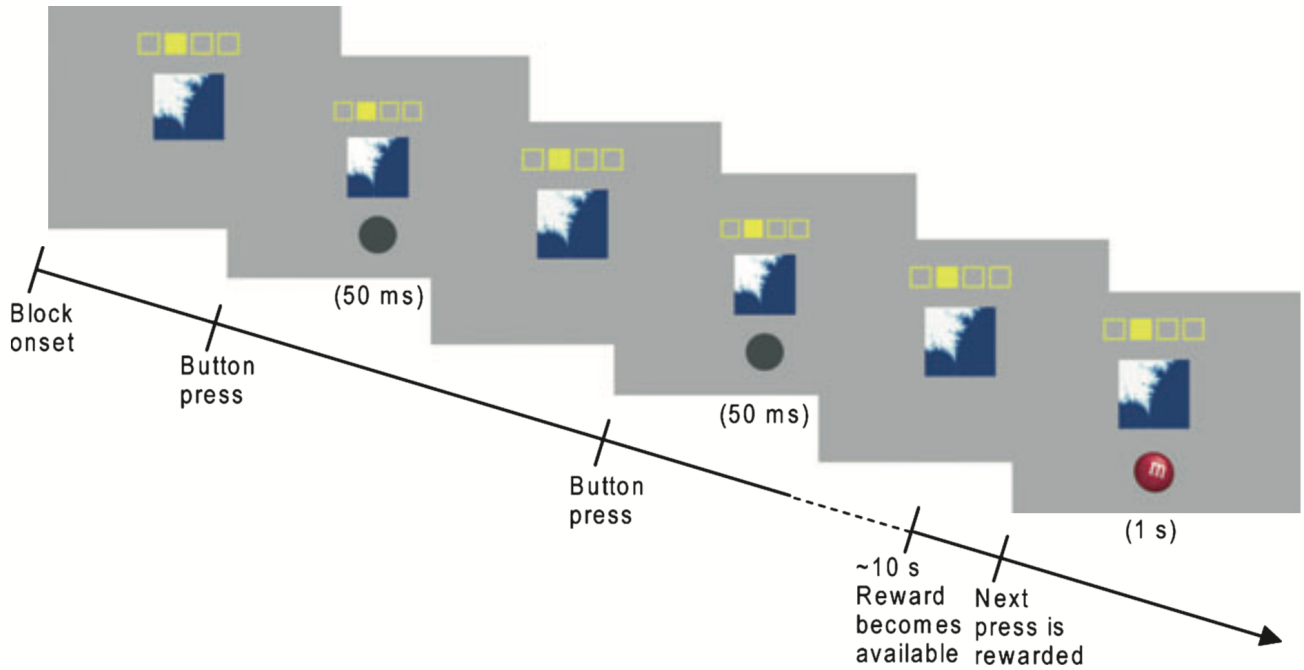
Model-based system



Model-free sys.

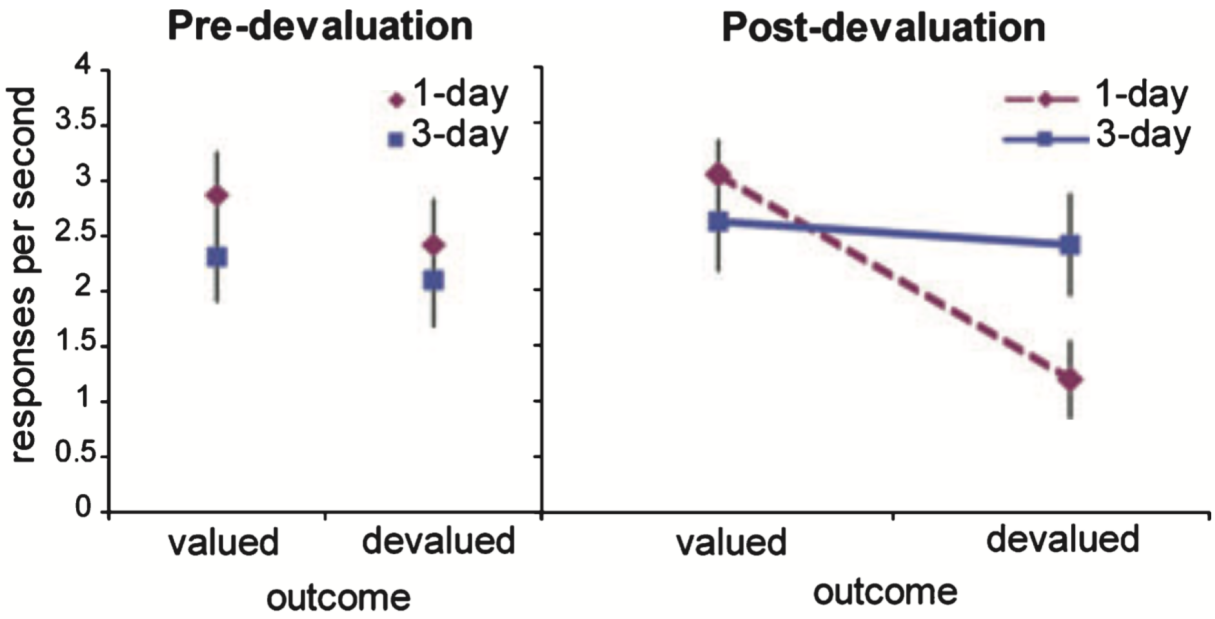
Behavior is initially model-based (goal-directed) and becomes model-free (habitual) with overtraining (Daw et al., 2005).

Habit learning in humans



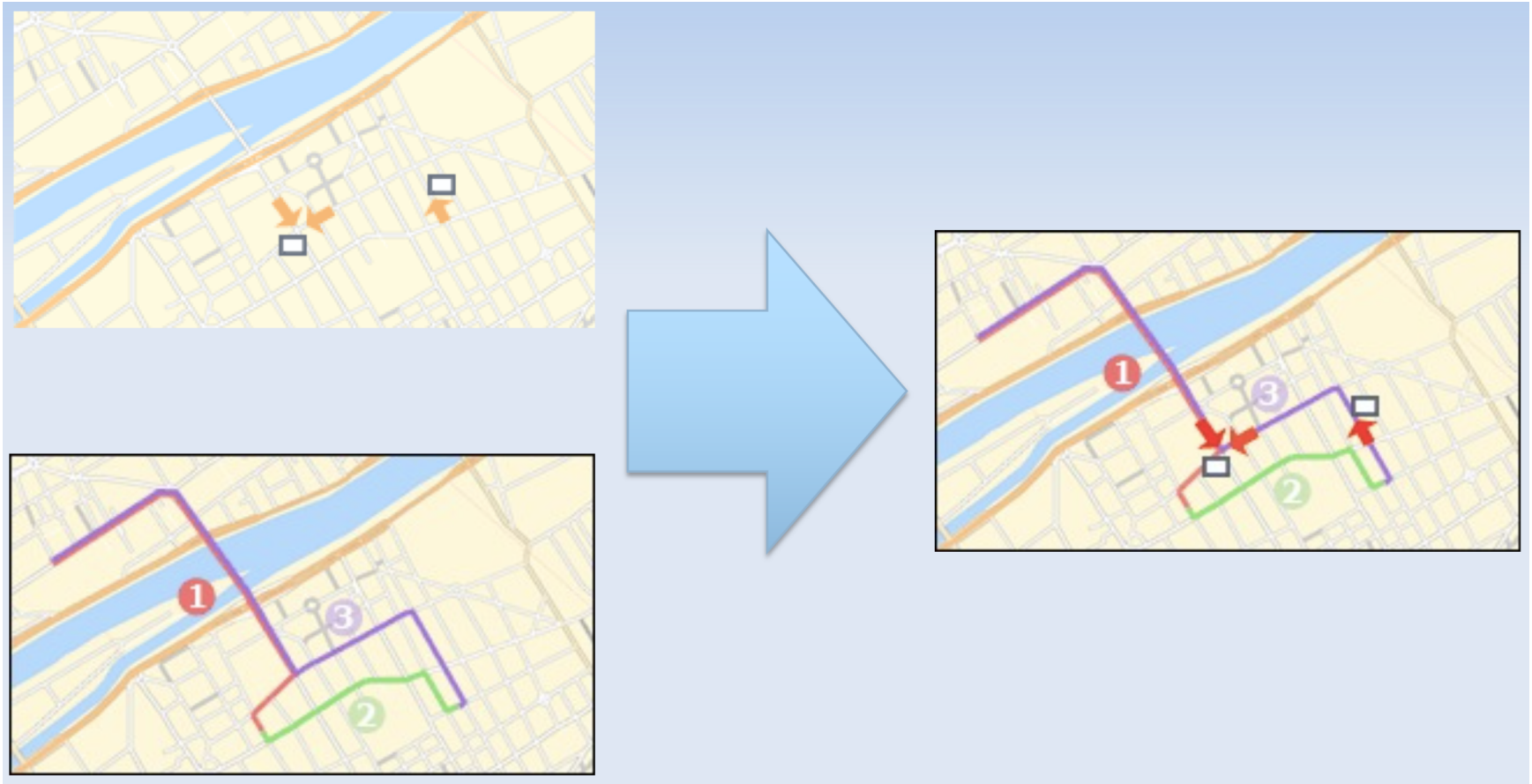
Tricomi Balleine O'Doherty 2009 EJN
One button is associated to M&M's, another button to Fritos.
Variable Interval (VI) schedule.

Habit learning in humans



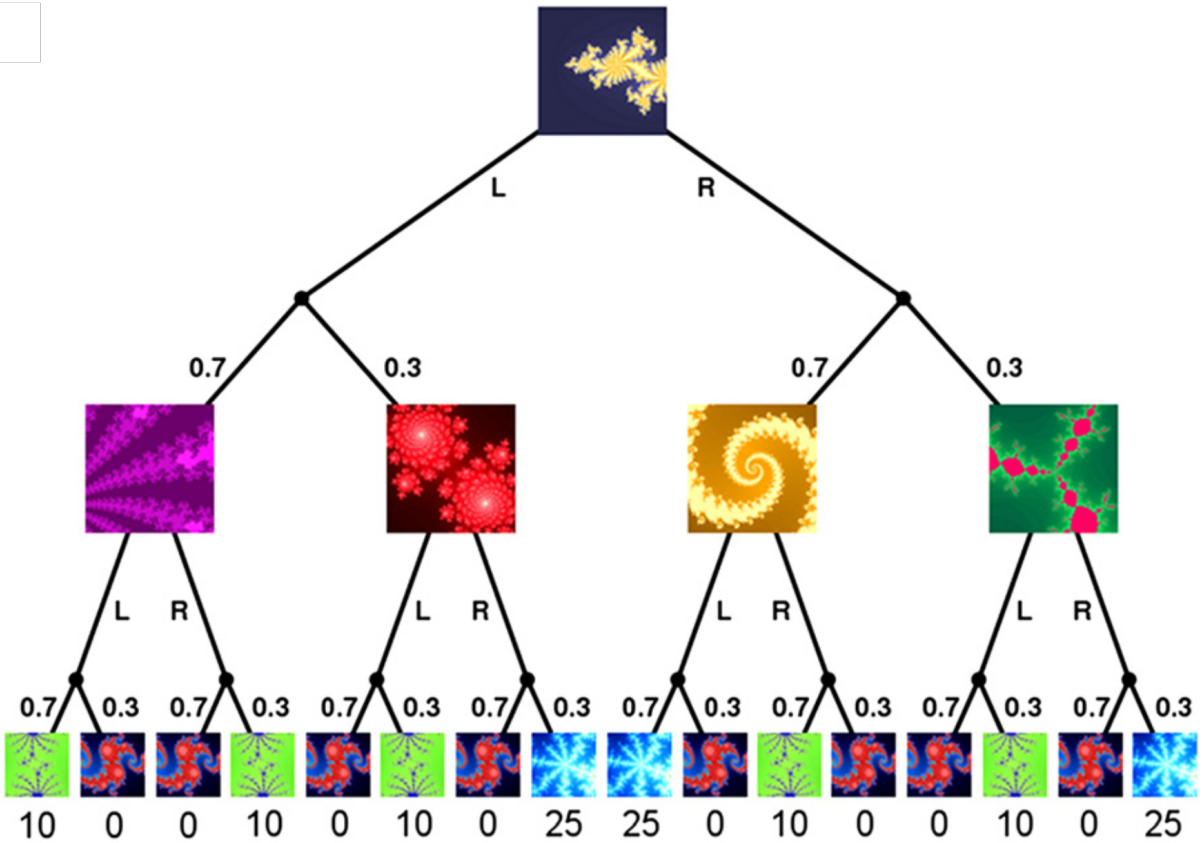
Tricomi Balleine O'Doherty 2009 EJN
Two groups (1-day training; 2 sessions vs. 3-day training; 12 sessions).
Outcome devaluation (selective satiation) of one of the outcomes.
The 3-day group (overtrained) continues to press after outcome devaluation.

Habitual navigation after long exposure



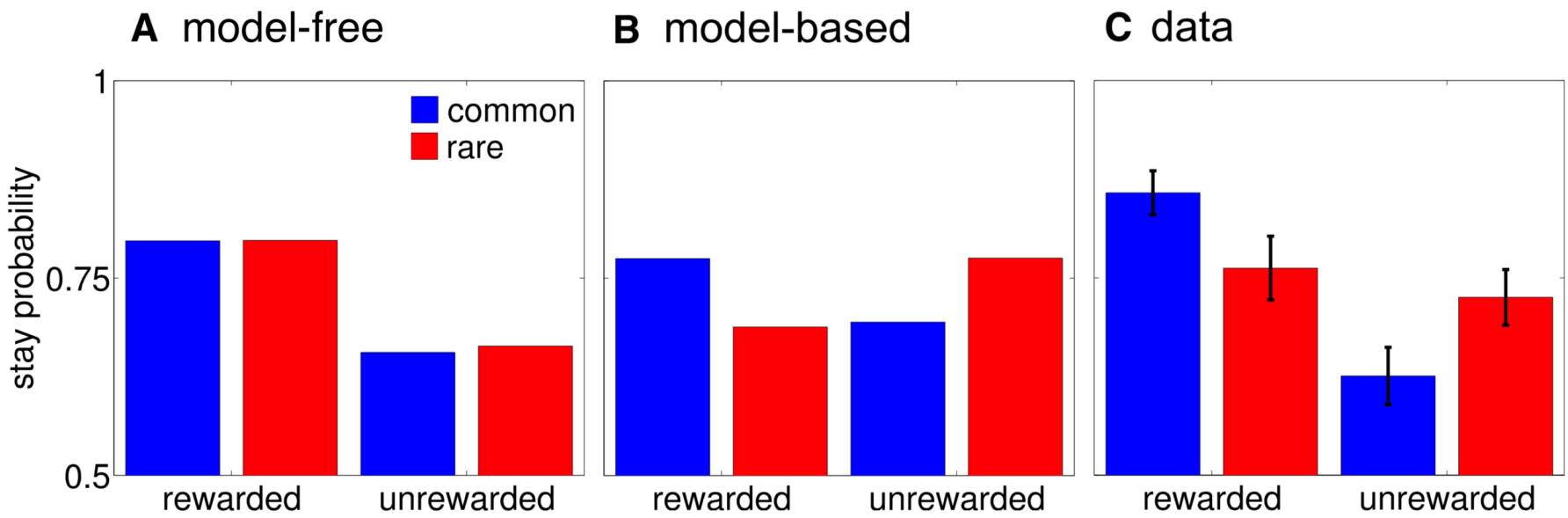
Khamassi & Humphries (2012) Frontiers in Behavioral Neuroscience

Example of model-based prospective search: The two-step task in humans



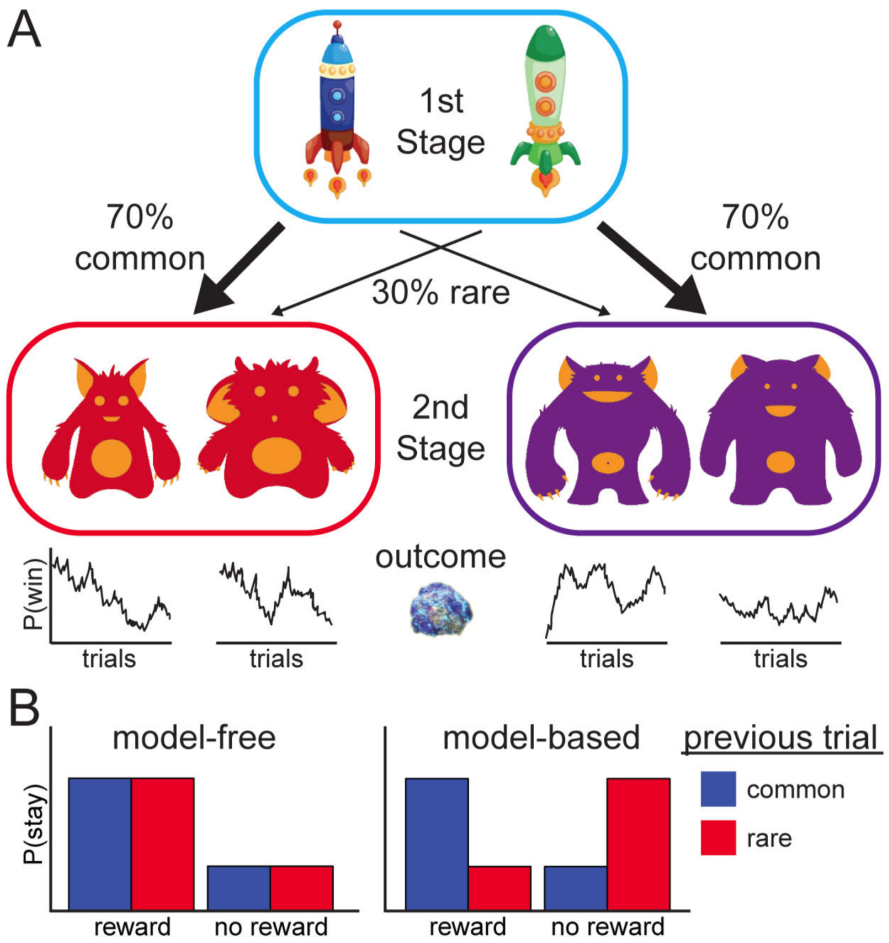
Glascher et al. (2010) Neuron; Daw et al. (2011) Neuron

Example of model-based prospective search: The two-step task in humans



Adult humans' behavior looks like a mixture of MF and MB.
Gläscher et al. (2010) Neuron; Daw et al. (2011) Neuron

The two-step task in children and teenagers



Decker et al. (2016) Psychological Science

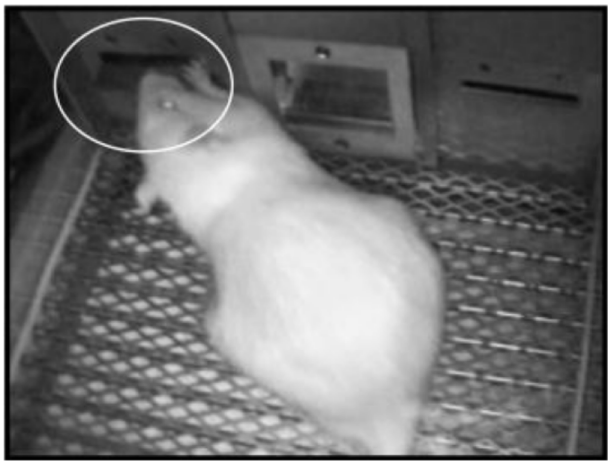
The two-step task in children and teenagers

Children relie less on MB and more on MF than adults.

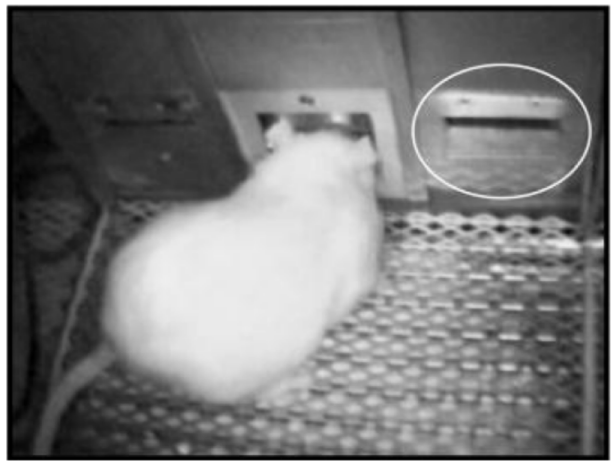
Decker et al. (2016) Psychological Science

Individual differences in MB/MF learning

Sign-trackers **MF**



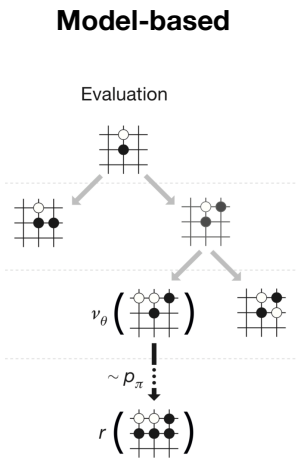
Goal-trackers **MB**

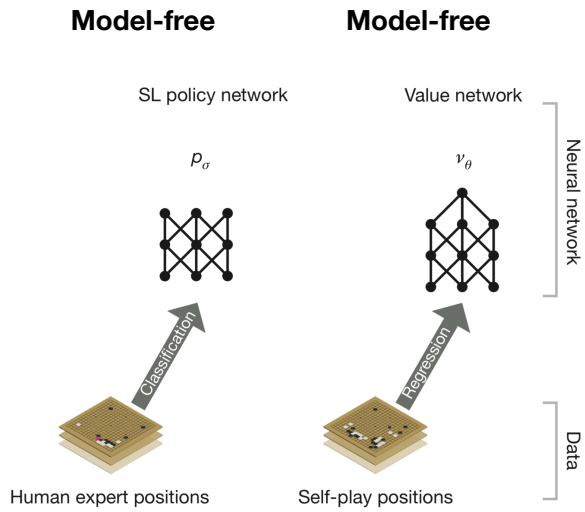


Collaboration with Flagel/Robinson (NIH Ann-Arbor), Coutureau/Marchand (CNRS Bordeaux) & Schoenbaum/Roesch (NIH / Univ. Maryland), Project NSF-NIH-ANR CRCNS :

Lesaint .. Khamassi (2014) PLoS Computational Biology
Lee, Gentry .. Khamassi, Roesch (2018) PLoS Biology
Cinotti .. Khamassi (2019) Nature Scientific Reports

Application: AlphaGo from Google Deepmind





The model-based system performs tree-search, while the model-free system learns "intuitions" like professional players.

Silver et al. (2016) Nature

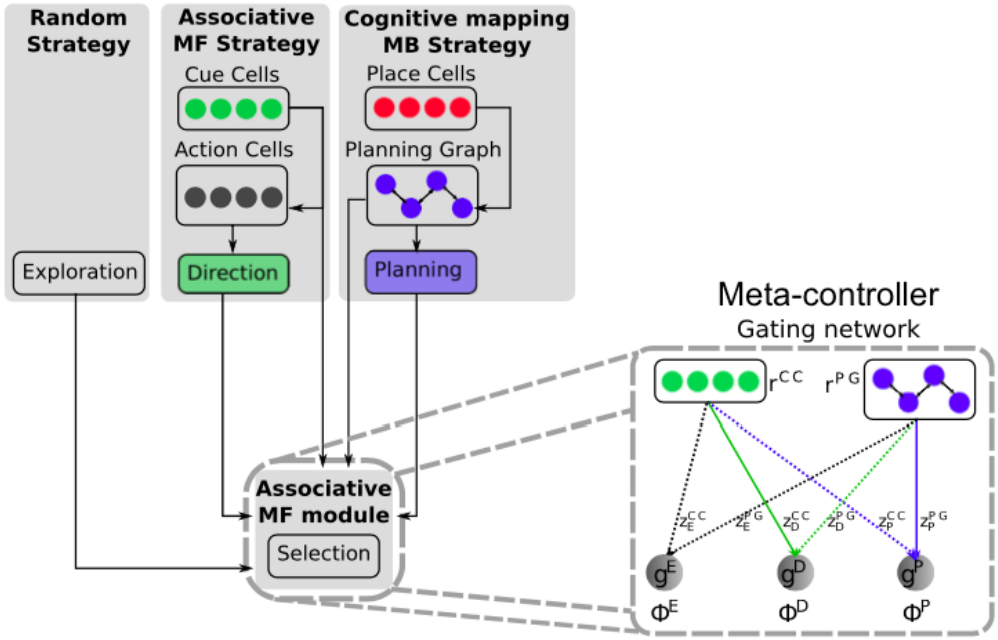
MB/MF coordination (I)

Computational models combining model-based (MB) and model-free (MF) reinforcement learning (RL)

- **Guazzelli Bota Corbacho Arbib (1998)**: World graph theory + affordances for frog navigation.
- **Daw Niv Dayan (2005)**: MB + MF for rat instrumental conditioning.
- **Keramati et al. (2011), Pezzulo et al. (2013), Lesaint et al. (2014)**: idem.
- **Collins & Frank (2012), Viejo et al. (2015, 2018)**: primates.
- **Khamassi & Humphries (2012)**: re-classifying navigation strategies with the MB/MF-RL framework.

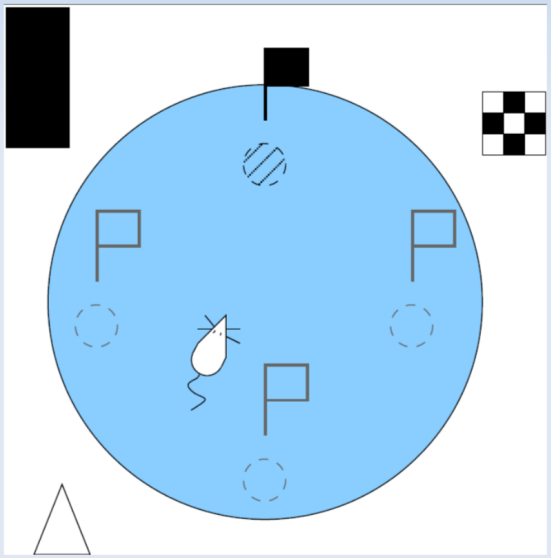
Open question: Which mechanism for MB/MF coordination/arbitration? Uncertainty? Average reward? Cooperation? Meta-learning?

MB/MF coordination (II)

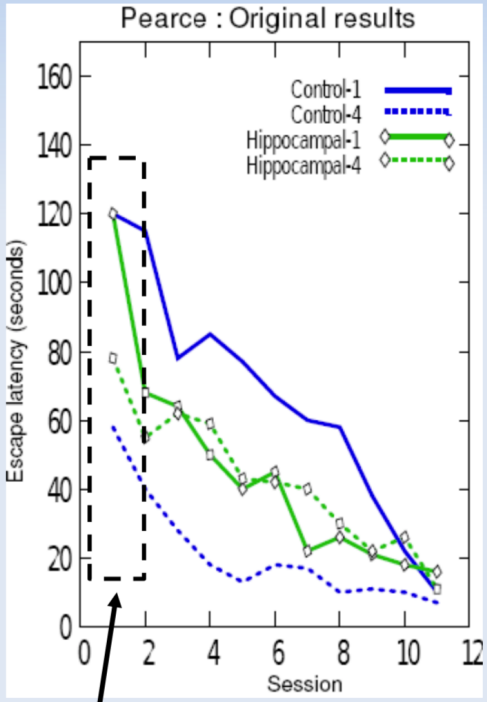


- [Dollé Chavarriaga Guillot Khamassi (2018) PLoS Computational Biology]: medial prefrontal cortex (mPFC) **meta-controller learning when to choose model-based or model-free RL** (applied to rat navigation). Similar in spirit to [Holroyd & McClure 2015] mPFC as hierarchical learner.
- [Coutureau & Killcross 2003, Killcross & Coutureau 2003] Consistent with lesion studies of rat mPFC impairing shift between learning systems but not the systems themselves.

Task with a cued platform (visible flag) changing location every 4 trials



Task of Pearce et al., 1998

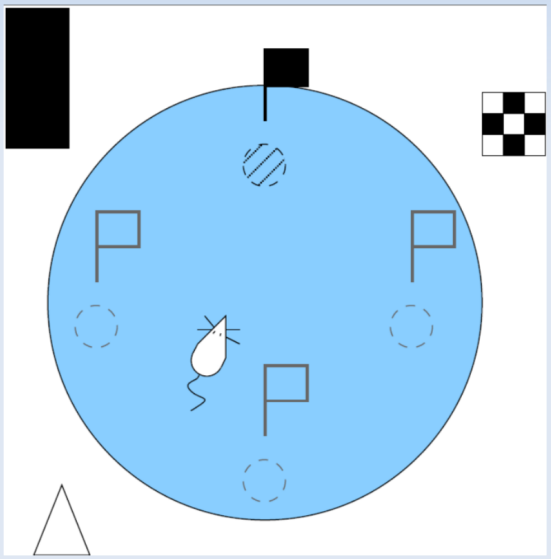


Rapid adaptation between trail #1 and trial #4

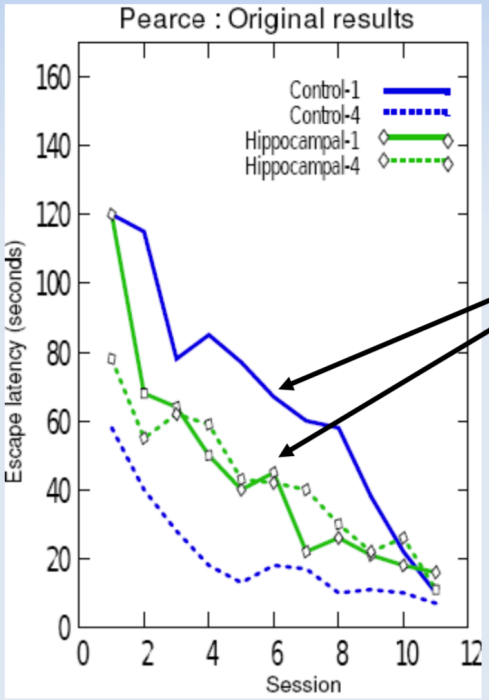
Not possible with a hippocampal lesion

[Dollé Chavarriaga Guillot Khamassi (2018) PLoS Computational Biology]

Task with a cued platform (visible flag) changing location every 4 trials



Task of Pearce et al., 1998

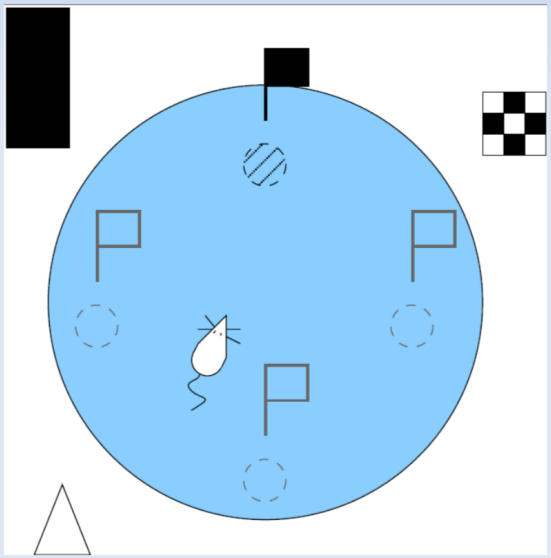


Hip-lesioned rats are better than controls at trial #1,

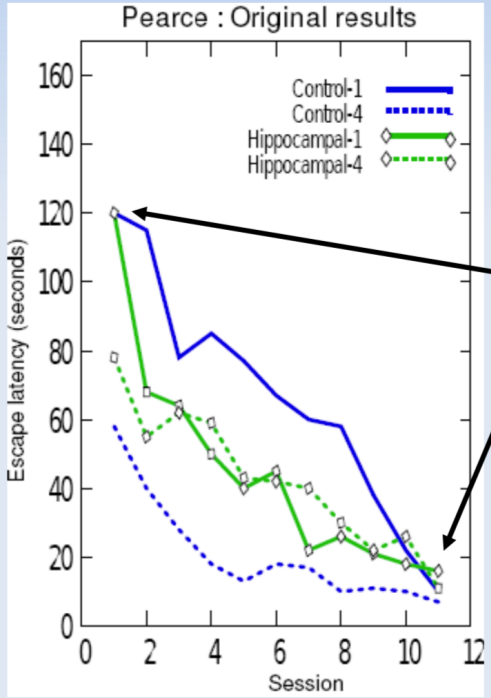
because the hippocampus-based strategy leads rats to the previous location of the platform.

[Dollé Chavarriaga Guillot Khamassi (2018) PLoS Computational Biology]

Task with a cued platform (visible flag) changing location every 4 trials



Task of Pearce et al., 1998

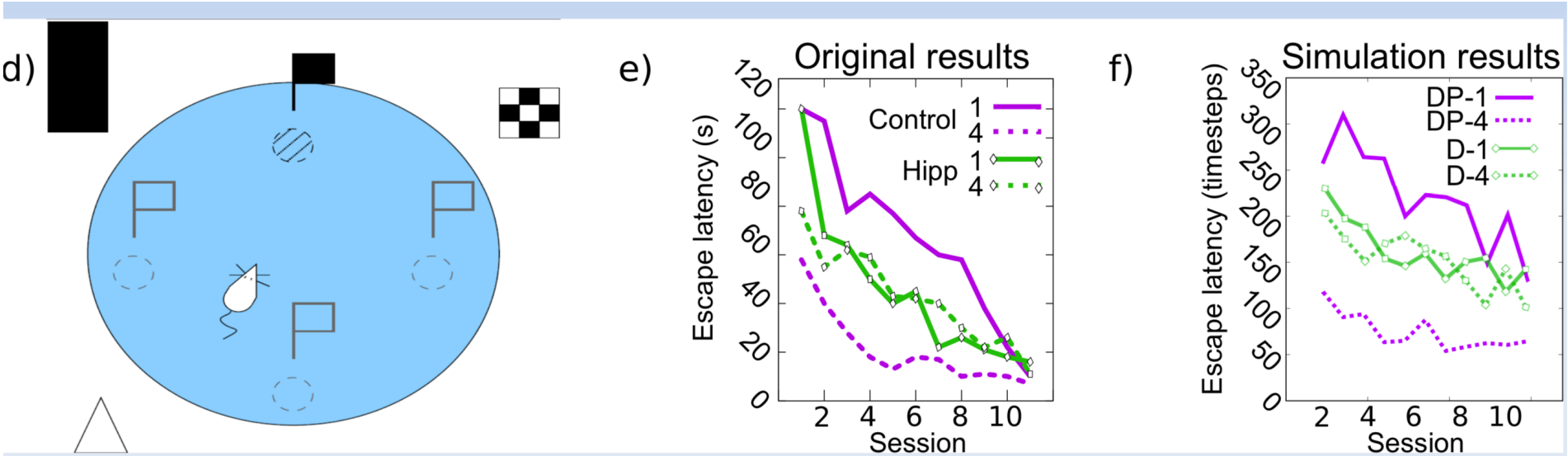


Progressive transfer from a hip-dependent place-based strategy to a cue-guided strategy:

Rats no longer loose time at the previous location of the platform.

[Dollé Chavarriaga Guillot Khamassi (2018) PLoS Computational Biology]

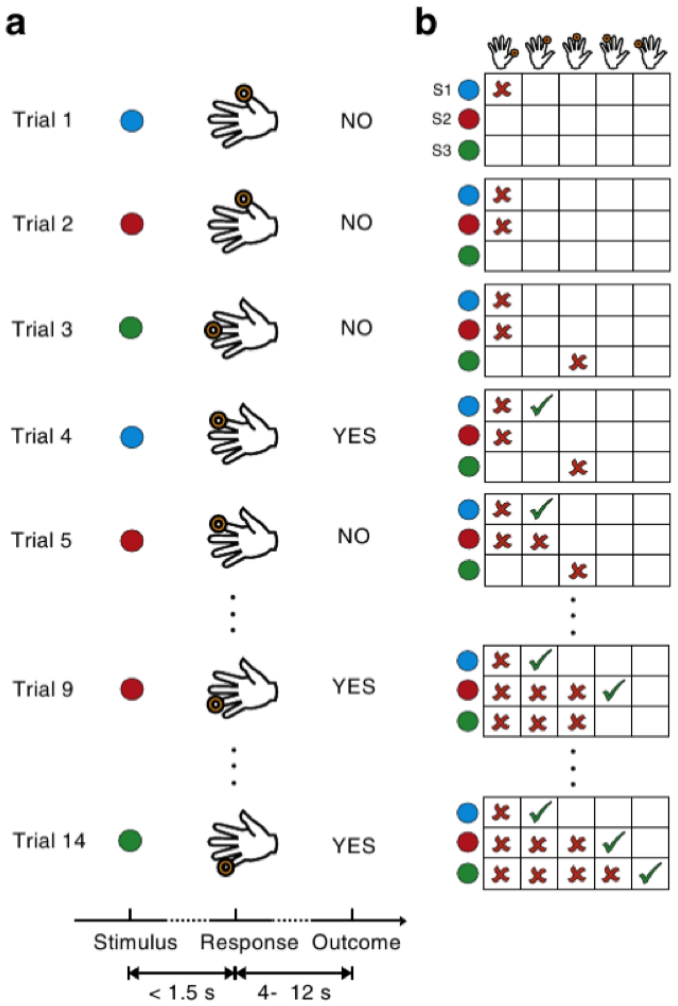
Task with a cued platform (visible flag) changing location every 4 trials



Task of Pearce et al., 1998 Nature

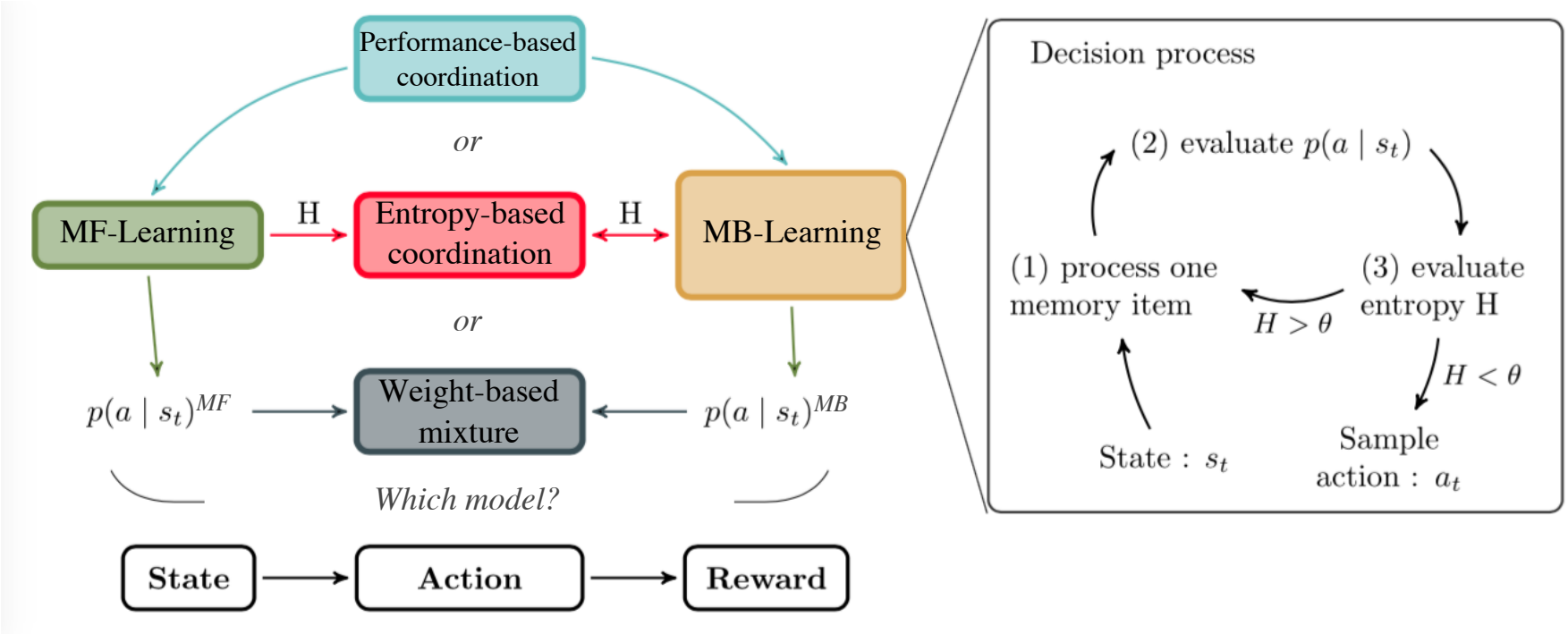
[Dollé Chavarriaga Guillot Khamassi (2018) PLoS Computational Biology]

Studying the coordination of MB and MF systems in humans



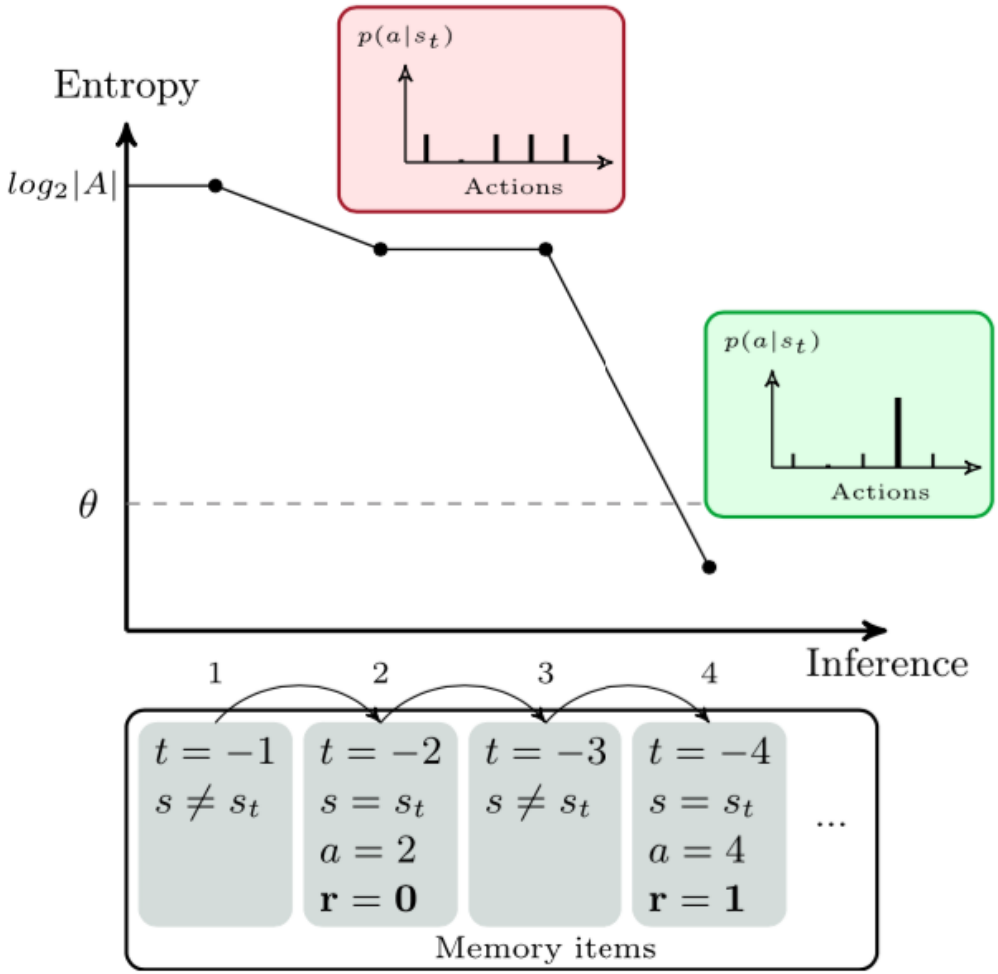
- Collaboration with Andrea Brovelli (CNRS Marseille)
- 4 blocks of trials
- 3 stim (blue, red, green)
- 5 options (fingers)
- Viejo et al (2015) Frontiers in Behavioral Neuroscience

Tested computational models



Viejo et al. (2015) Frontiers in Behavioral Neuroscience

Tested computational models



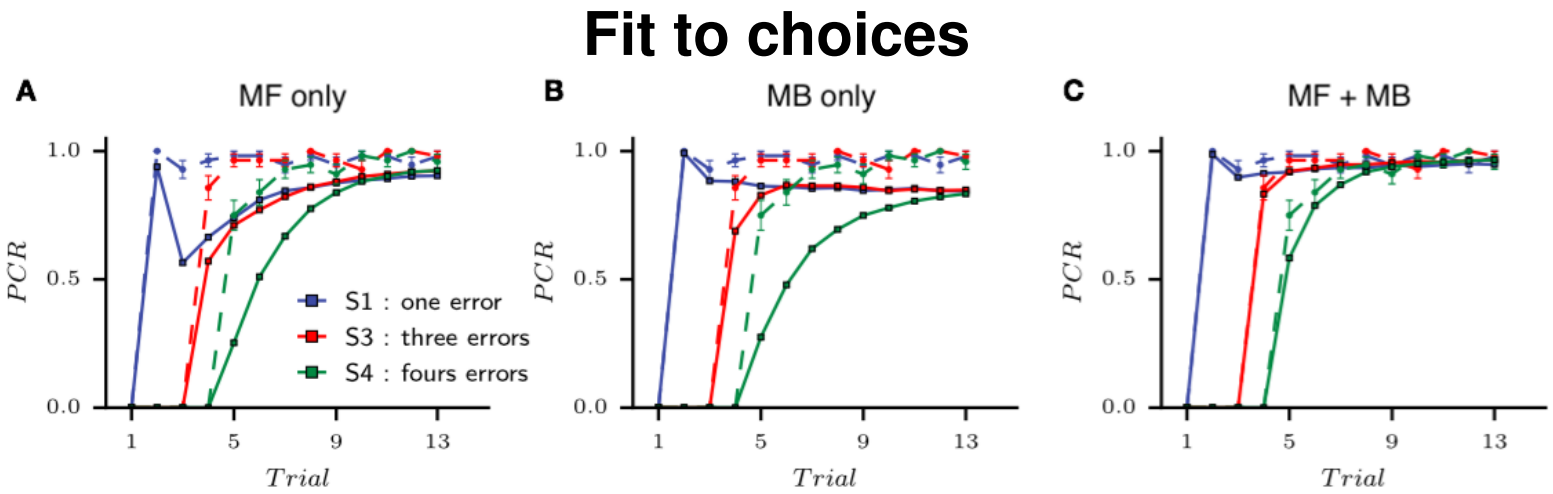
Adaptive working-memory with a subject-specific entropy threshold (θ) and a memory decay parameter (ϵ).

Model comparison results

Subject	-Bloc 1	-Bloc 2	-Bloc 3	-Bloc 4	All blocs
1	**E-Coord**	W-Mix	**E-Coord**	**E-Coord**	W-Mix
2	**VPI-select**	E-Coord	E-Coord	E-Coord	E-Coord
3	E-Coord	E-Coord	E-Coord	E-Coord	E-Coord
4	E-Coord	E-Coord	E-Coord	E-Coord	E-Coord
5	**W-Mix**	E-Coord	E-Coord	E-Coord	E-Coord
6	E-Coord	E-Coord	E-Coord	**W-Mix**	E-Coord
7	**E-Coord**	**VPI-select**	**VPI-select**	W-Mix	W-Mix
8	**W-Mix**	VPI-select	VPI-select	**E-Coord**	VPI-select
9	**VPI-select**	**VPI-select**	**VPI-select**	**VPI-select**	W-Mix
10	VPI-select	VPI-select	VPI-select	VPI-select	VPI-select
11	E-Coord	E-Coord	E-Coord	E-Coord	E-Coord
12	E-Coord	**W-Mix**	E-Coord	**W-Mix**	E-Coord
13	E-Coord	E-Coord	E-Coord	E-Coord	E-Coord
14	E-Coord	E-Coord	E-Coord	E-Coord	E-Coord

Viejo et al. (2015) Frontiers in Behavioral Neuroscience

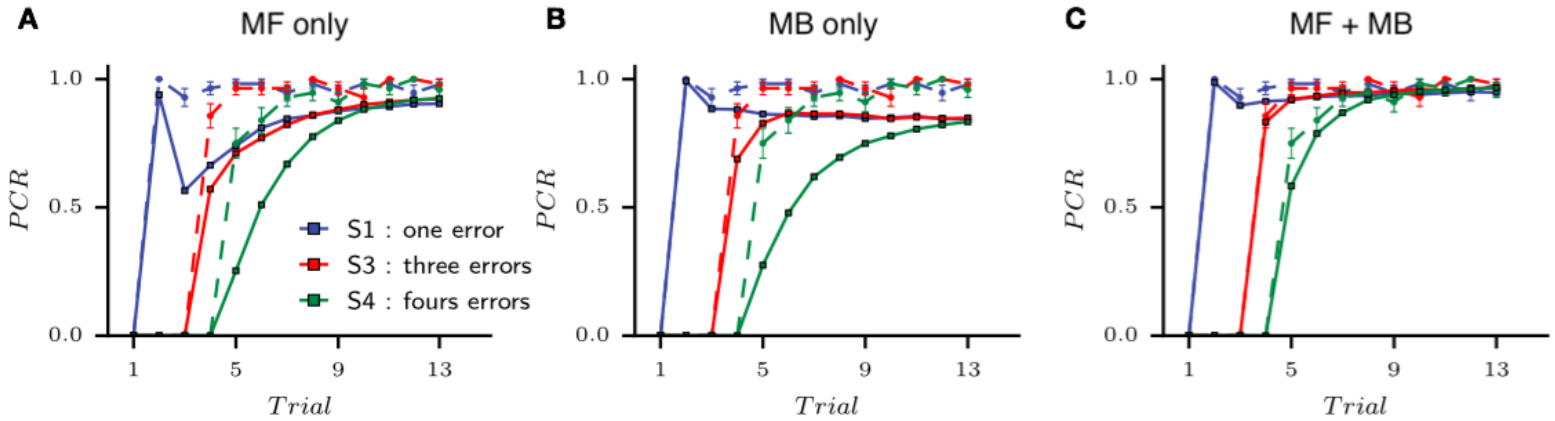
Model fitting results



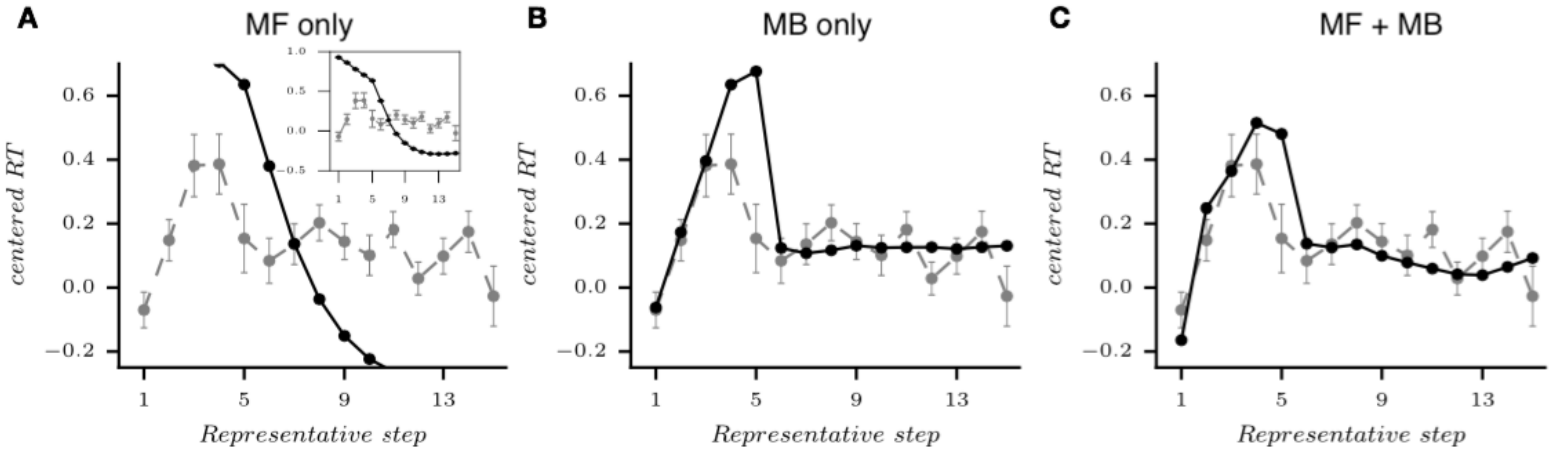
Viejo et al. (2015) Frontiers in Behavioral Neuroscience

Model fitting results

Fit to choices

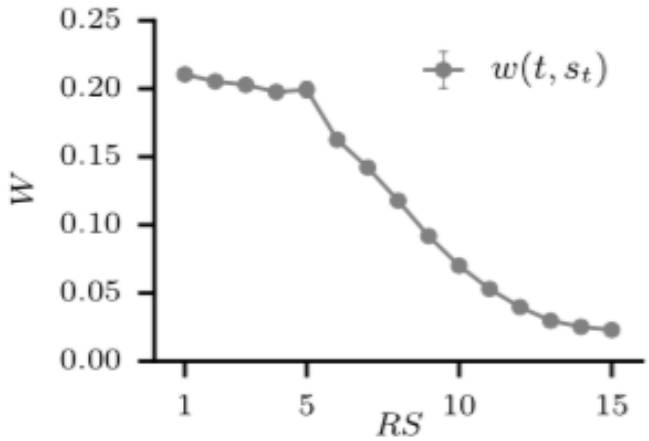


Fit to reaction times



Viejo et al. (2015) Frontiers in Behavioral Neuroscience

Trial-by-trial contribution of the MB system to the subjects' decisions according to the optimized model



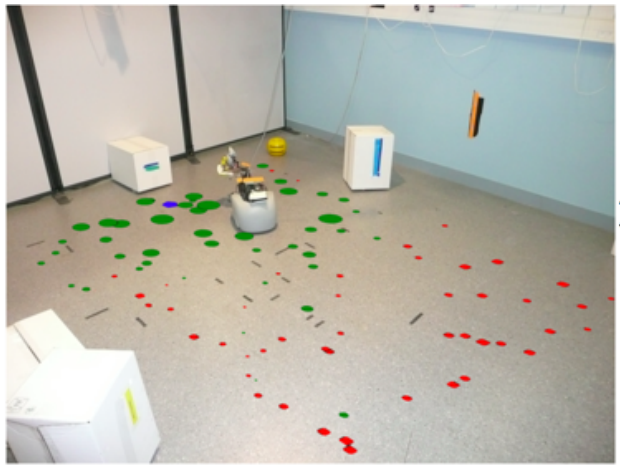
Viejo et al. (2015) Frontiers in Behavioral Neuroscience

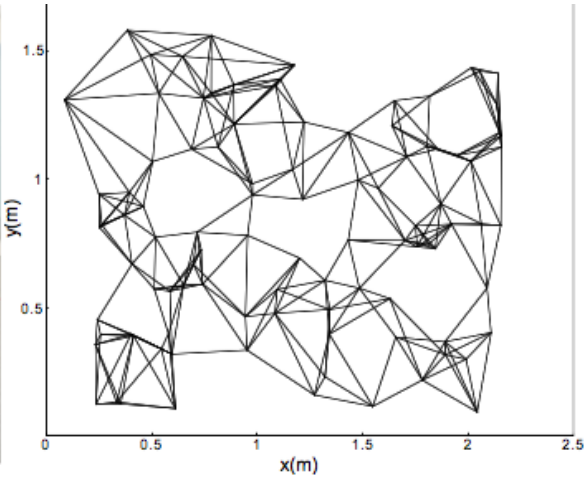
Applications of MB/MF coordination to Robotics

Robotics experimentation

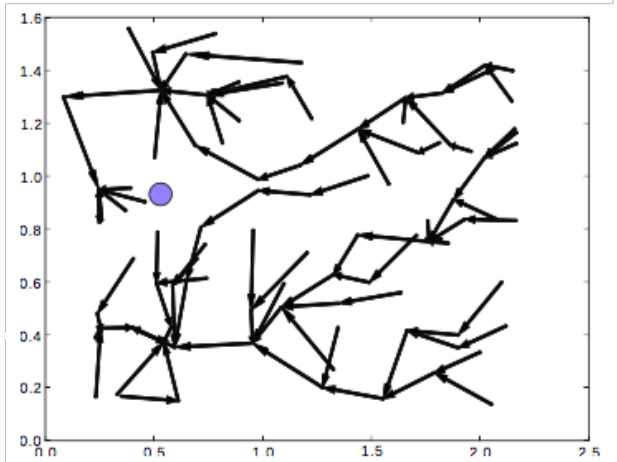
[Meyer et al. 2005, Caluwaerts et al. 2012]: Navigation experiments with the Psikharpax robot. National CNRS Project ROBEA, EU FP6 Project ICEA.

Robotics experimentation

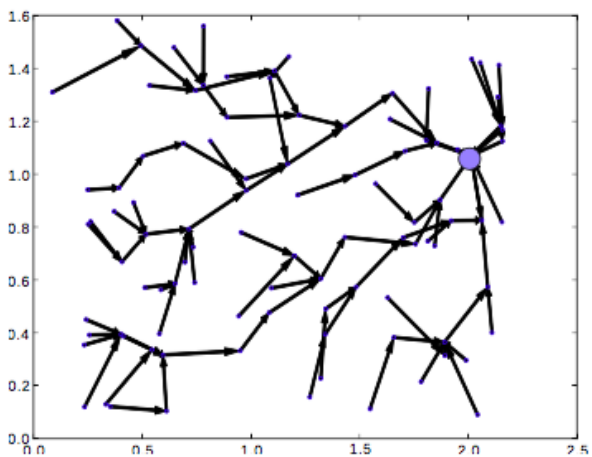




(b) Topological map constructed by the robot.



(c) Paths towards the goal



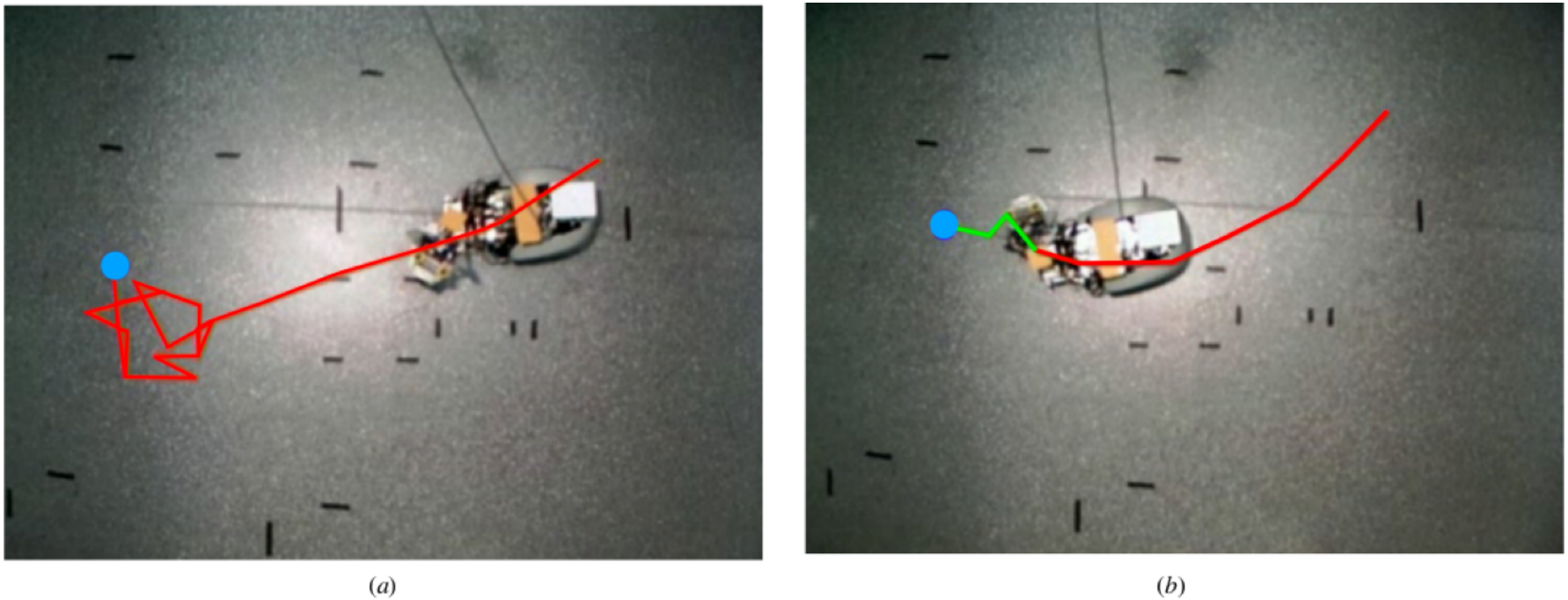
(d) Paths towards the goal

[Caluwaerts et al. 2012]: Robot MB learning.

Robotics experimentation

MB strategy only

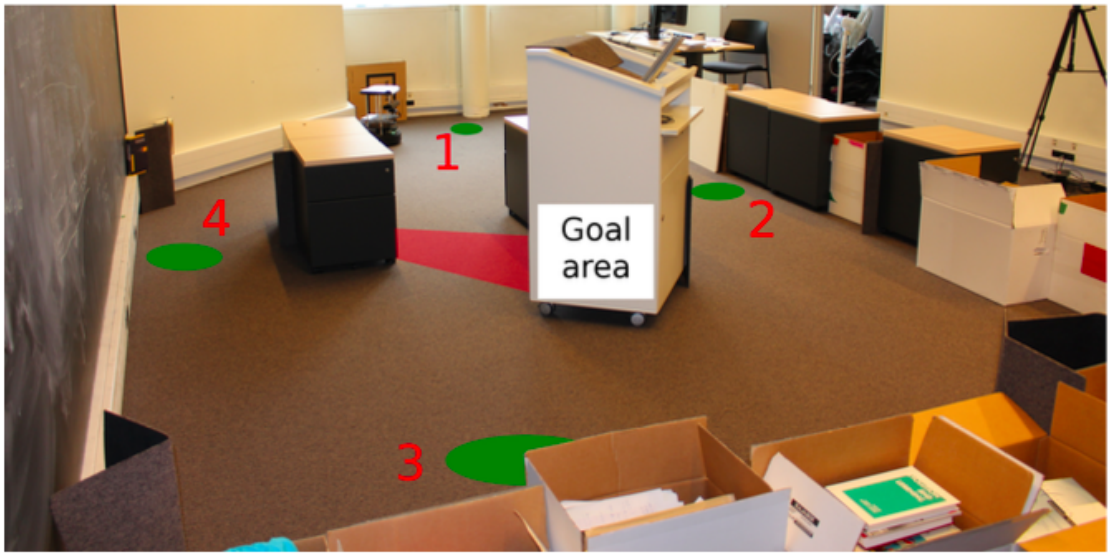
MB+MF strategies

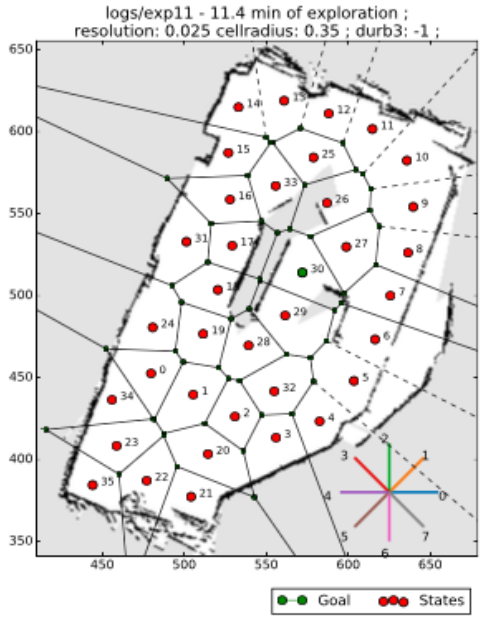


(a) (b)

[Caluwaerts et al. 2012]: MB-MF cooperation within trials. Red: trajectory controlled by the MB system. Green: trajectory controlled by the MF system.

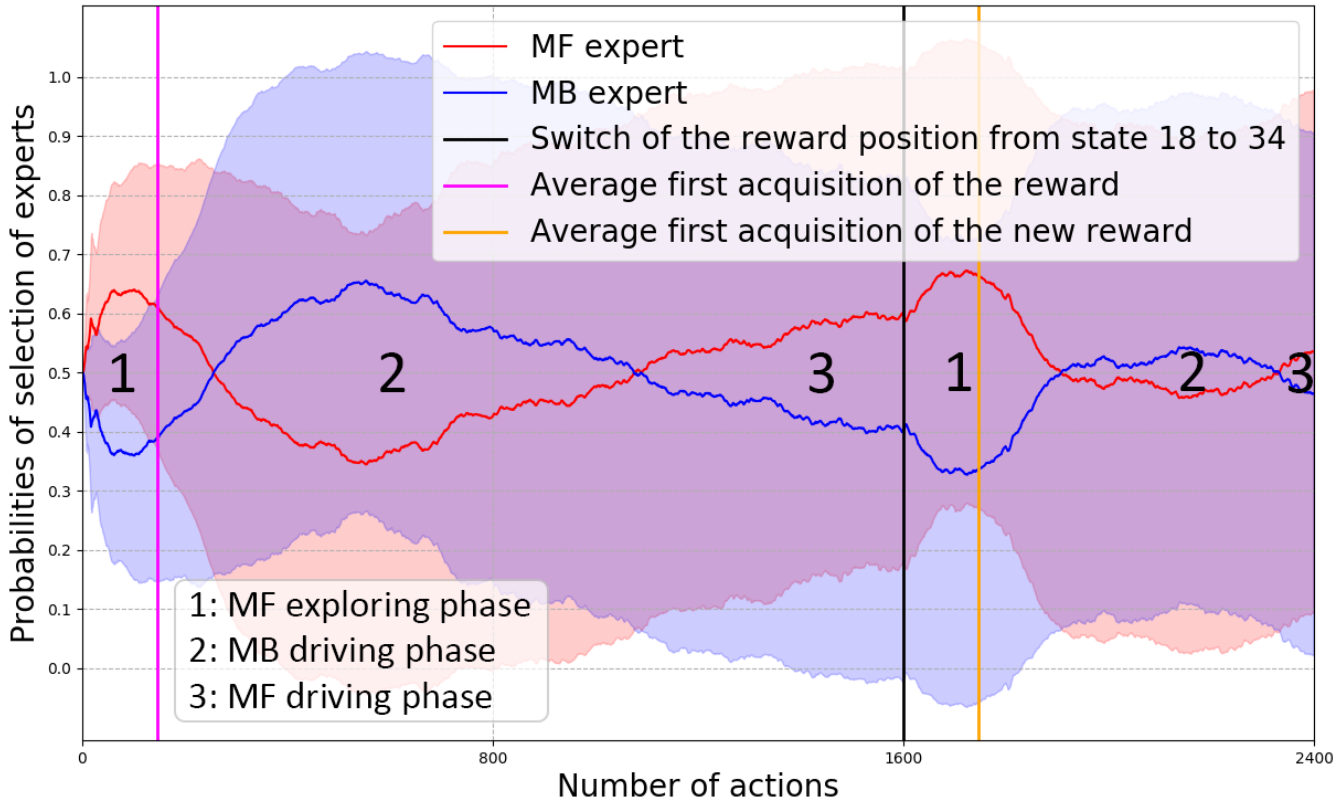
More recent robotics application





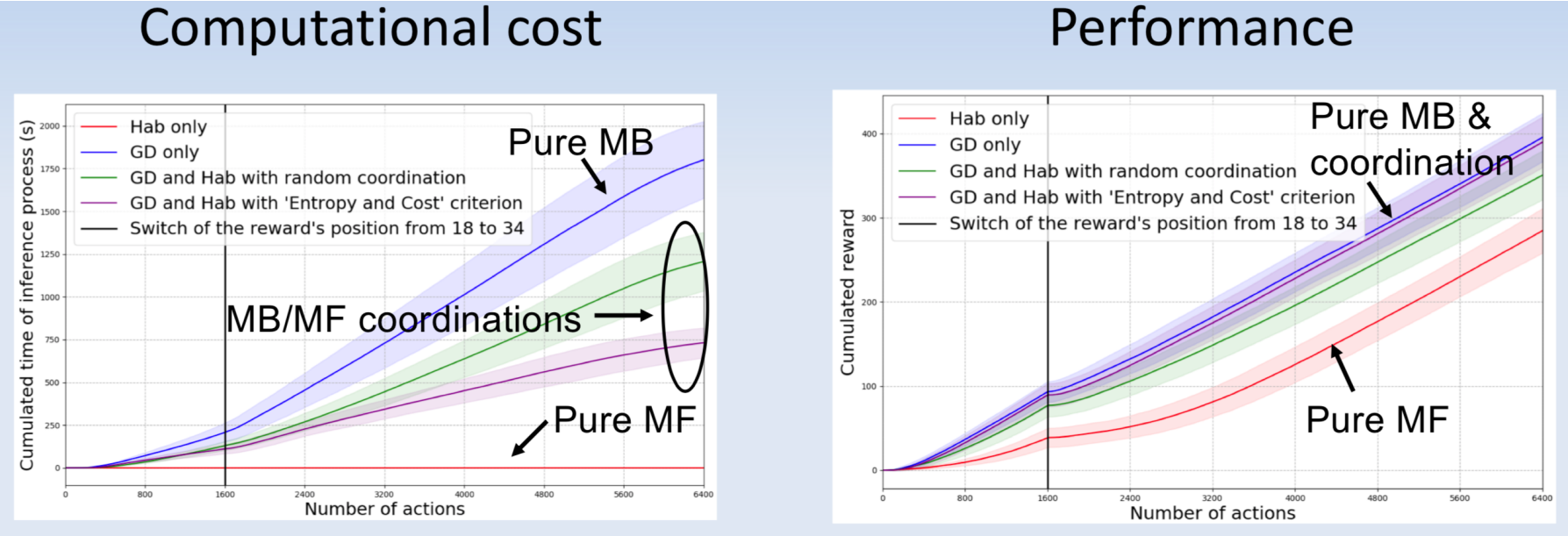
Dromnelle et al. (2020) Living Machines Conference

More recent robotics application



Dromnelle et al. (2020) Living Machines Conference

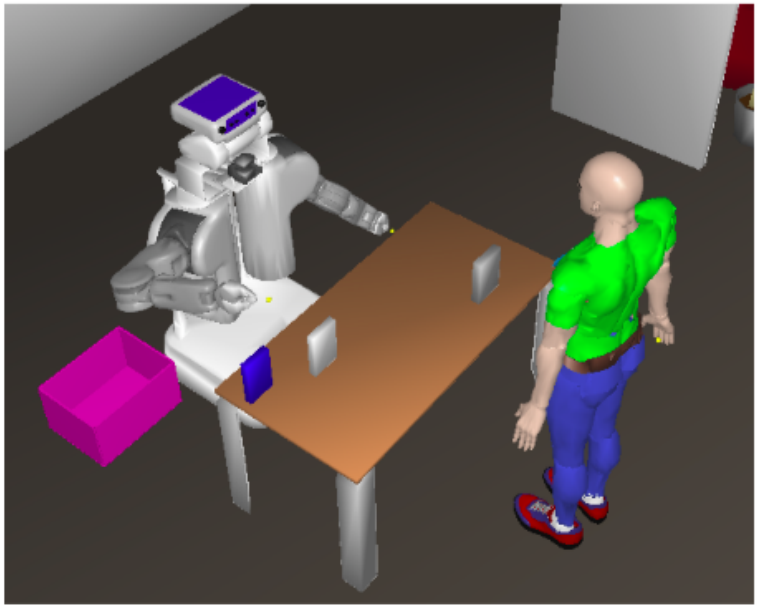
More recent robotics application

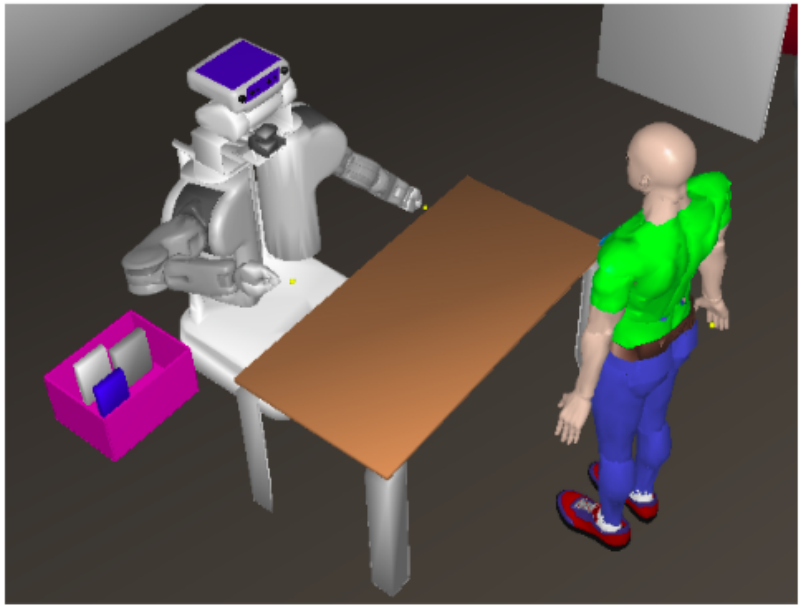


Dromnelle et al. (2020) Living Machines Conference

Prediction: MB/MF coordination should not only depend on uncertainty, but also on computational cost!

Robot habit learning

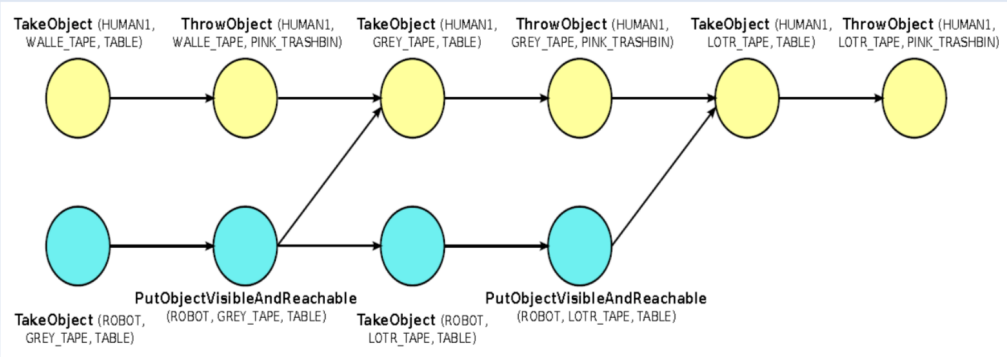




Task: Clean the table

Current state: A priori given action plan (right image)

Goal: Autonomous learning by the robot



Work of Erwan Renaudo in collaboration with CNRS-LAAS, Toulouse.

Application to child-robot interaction for assistive robotics

London **Athens**

Khamassi et al. (2018a) IEEE TCDS ; Zaraki*, Khamassi* et al. (2019)

Summary

Online adaptive coordination of model-based (MB) and model-free (MF) reinforcement learning (RL)

- The RL framework decomposes the world into states, actions, rewards, transitions.
- The RL agent tries to maximize the sum of future discounted rewards.
- MF-RL: slow learning, fast decision-making.
- MB-RL: fast learning, slow decision-making.
- The brain of mammals coordinates MB and MF RL mechanisms.
- Computational neuroscience models for MB-MF RL coordination.
- Applications to robotics for efficiency and computational cost reduction.

Acknowledgments

Collaborators

- Benoît Girard, Olivier Sigaud (Sorbonne)
- Mark Humphries (Nottingham, UK)
- Laurent Dollé (2010) (now in private sector)
- Ken Cauwaerts (2012) (now at NASA)
- Florian Lesaint (2014) (now in private sector)
- Guillaume Viejo (2016) (now post-doc in Canada)
- François Cinotti (2019) (now post-doc in the UK)

Open source

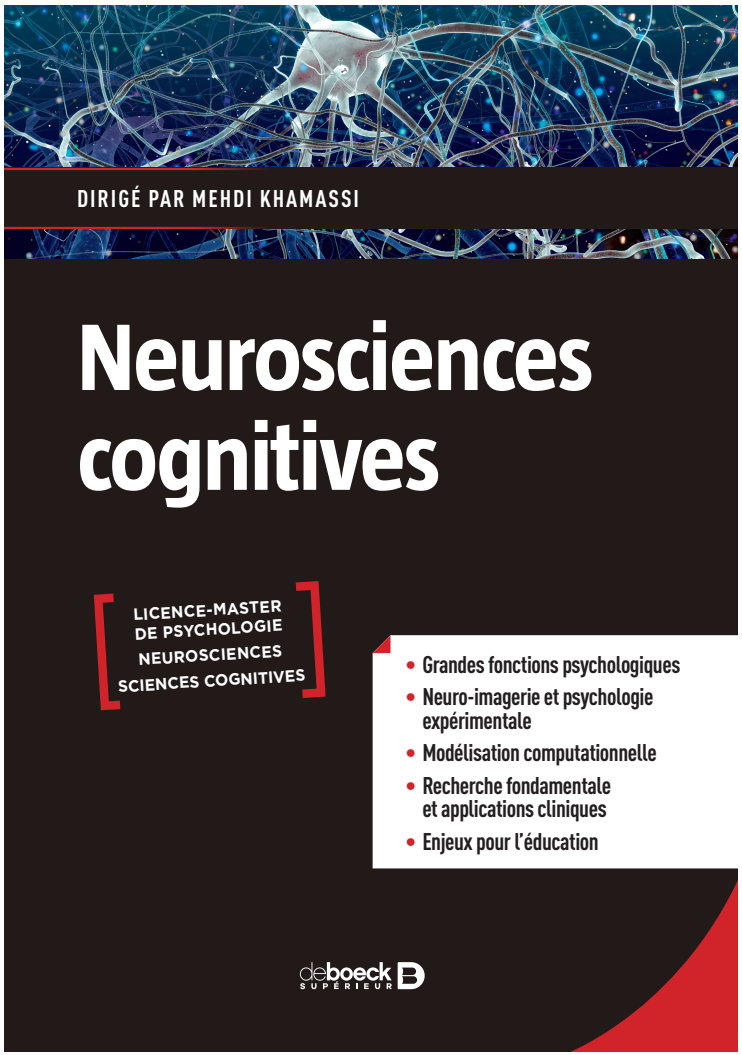
- https://github.com/MehdiKhamassi/RLwithReplay

Funding

- EU, ANR, CNRS, Sorbonne.

SUPPLEMENTARY MATERIAL

Khamassi (Ed.) (2021) Neurosciences Cognitives.



Chapitres

1 Perception et attention - Thérèse Collins et Laura Dugué
2 Le cerveau, le mouvement, et les espaces - Alain Berthoz
3 Étude des systèmes de mémoire dans le cadre d'un comportement : la navigation - Laure Rondi-Reig
4 Décision et action - Alizée Lopez-Persem et Mehdi Khamassi
5 Neurolinguistique - Perrine Brusini et Élodie Cauvet
6 Conscience et métacognition - Louise Goupil et Claire Sergent
7 Cognition sociale - Marwa El Zein, Louise Kirsch et Lou Safra
8 Psychologie et neurosciences : enjeux pour l'éducation - Emmanuel Sander et al.
9 Initiation à la modélisation computationnelle - Anne Collins et Mehdi Khamassi

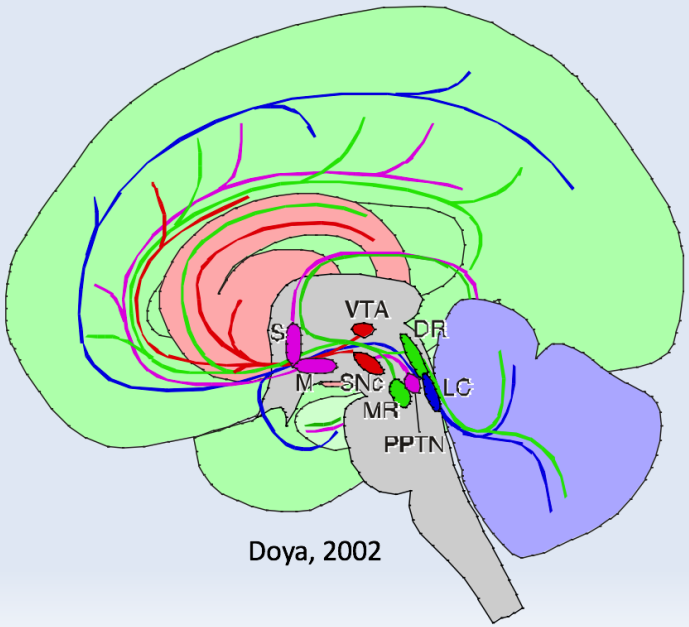
Meta-learning
(online parameter tuning)

Meta-learning in the brain

$$Q(s,a) \leftarrow Q(s,a) + \alpha \,.\, \delta \quad \longleftarrow \quad \textbf{Action values update}$$

$$\delta = r + \gamma \,.\, \max[Q(s',a')] - Q(s,a) \quad \longleftarrow \quad \textbf{Reinforcement signal}$$

$$P(a) = \frac{\exp(\beta \,.\, Q(s,a))}{\sum_b \exp(\beta \,.\, Q(s,b))} \quad \longleftarrow \quad \textbf{Action selection}$$



Doya, 2002

Dopamine: TD error δ

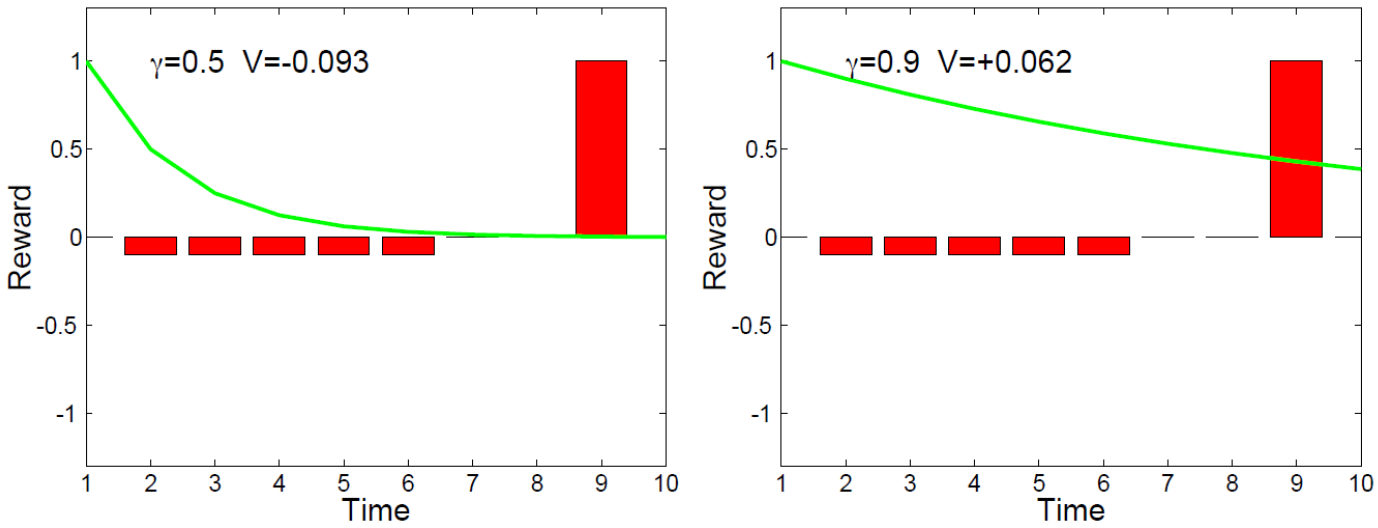
Acetylcholine: learning rate α

Noradrenaline: exploration β

Serotonin: temporal discount γ

Meta-learning in computational terms

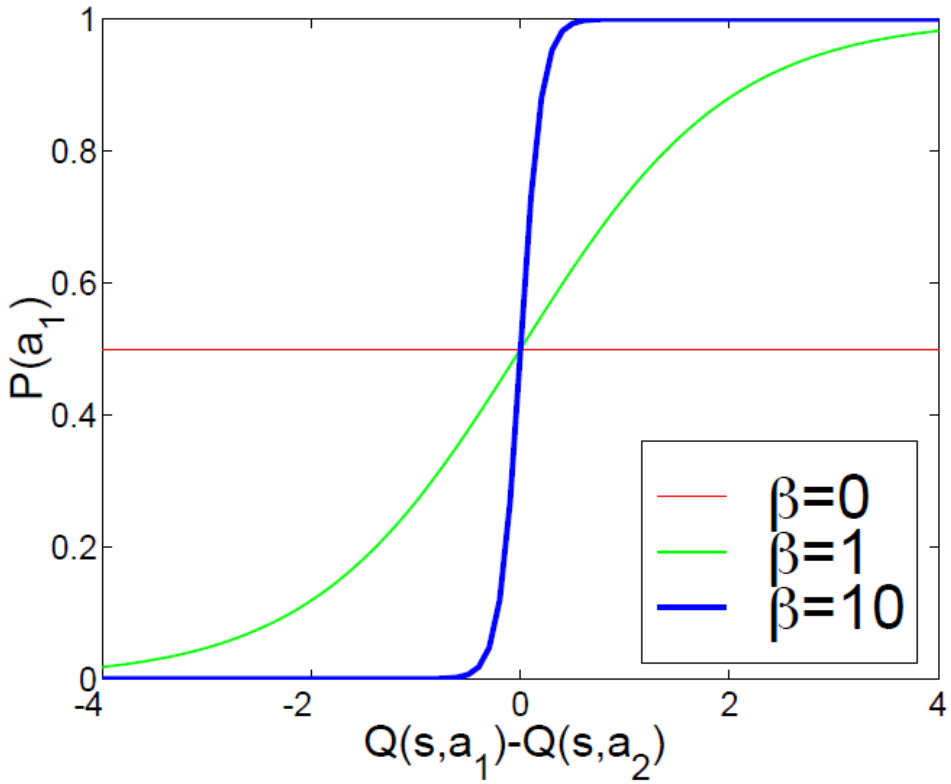
Effect of γ on expected reward value



Doya (2002)

Meta-learning in computational terms

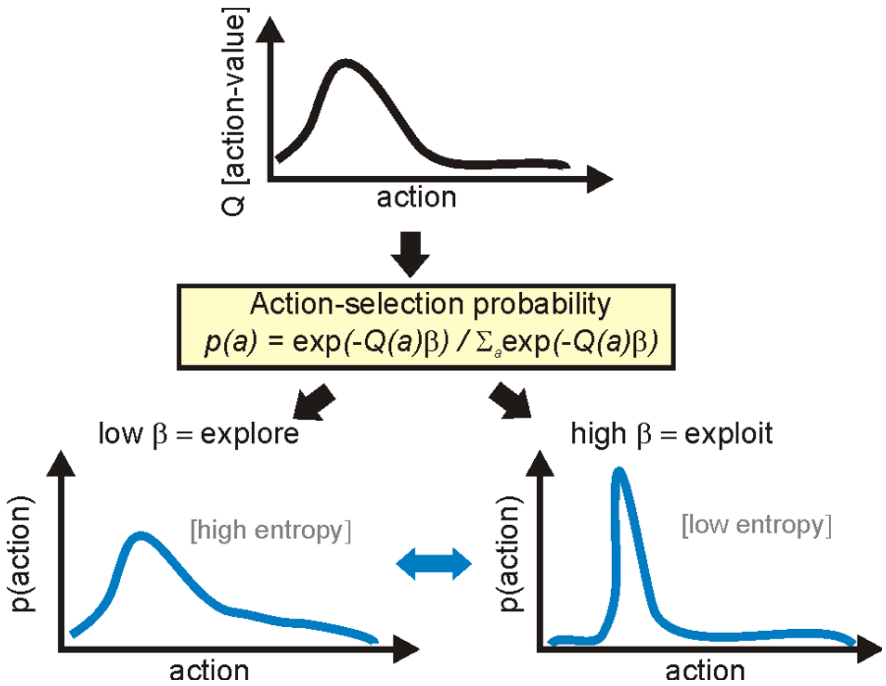
Effect of β on random exploration (softmax function)



Doya (2002)

Meta-learning in computational terms

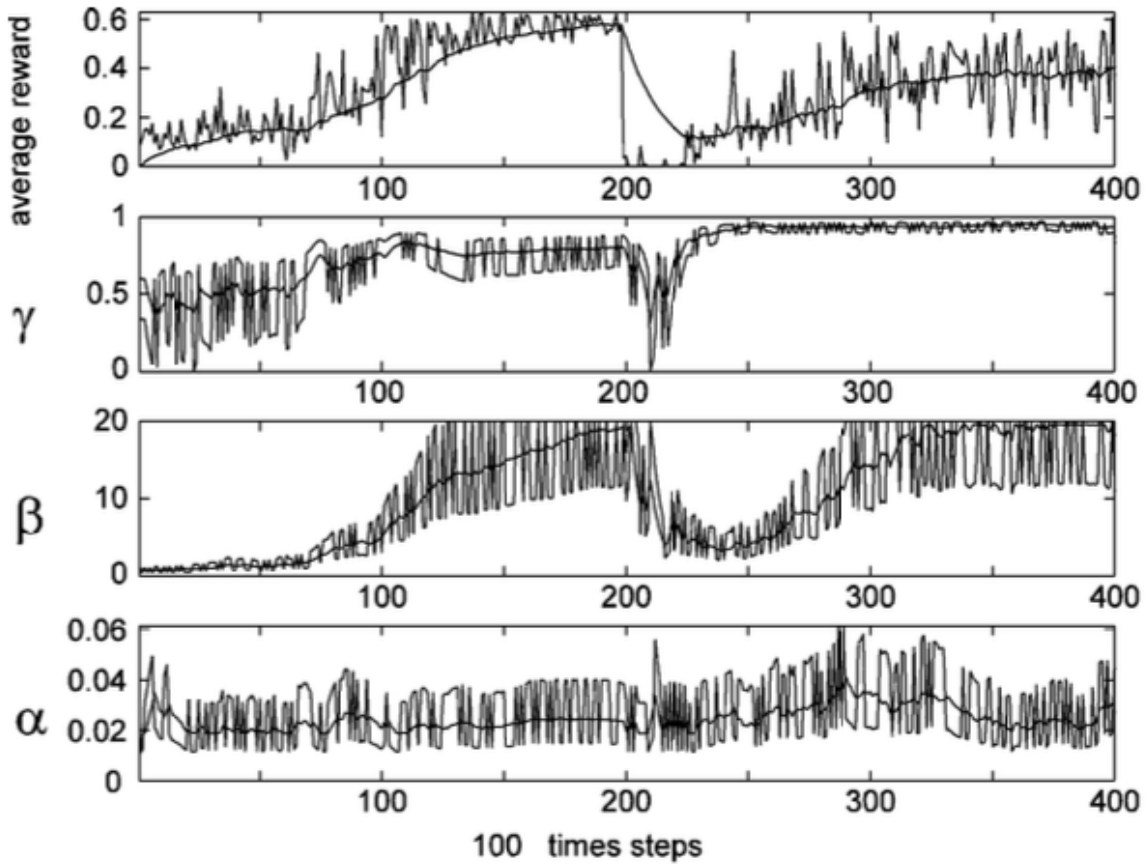
Exploration-exploitation trade-off: necessary for learning but affects action selection



Humphries, Khamassi & Gurney (2012)

Meta-learning in computational terms

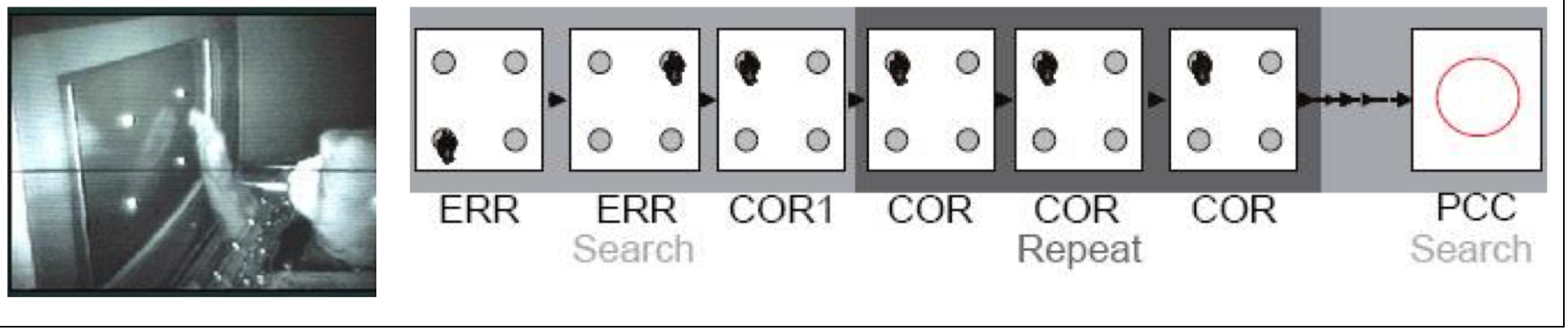
Tuning RL parameters as a function of average reward variations (volatility)



Schweighofer & Doya (2003)

Can we use such meta-learning principles to better understand neural mechanisms in the prefrontal cortex?

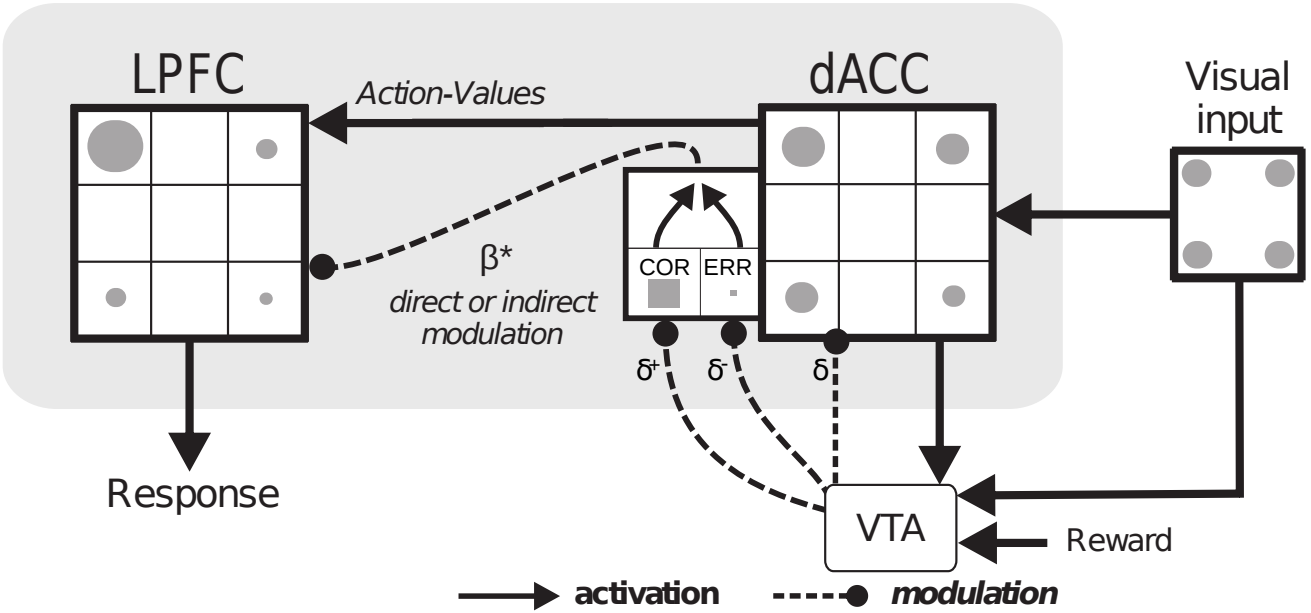
Meta-learning in neuroscience



Question: How did the monkeys learn to re-explore after each presentation of the PCC signal?
Hypothesis: By trial-and-error during pretraining.

Khamassi et al. (2011) Frontiers in Neurorobotics
Khamassi et al. (2013) Progress in Brain Research

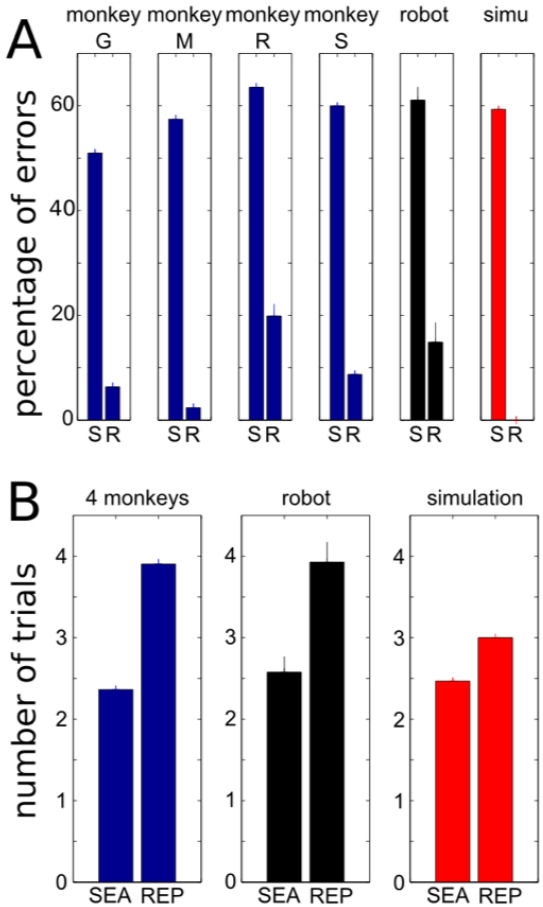
Meta-learning in neuroscience



β^*: exploratory variable computed from reward history and used to modulate β
Khamassi et al. (2011) Frontiers in Neurorobotics

Meta-learning in neuroscience

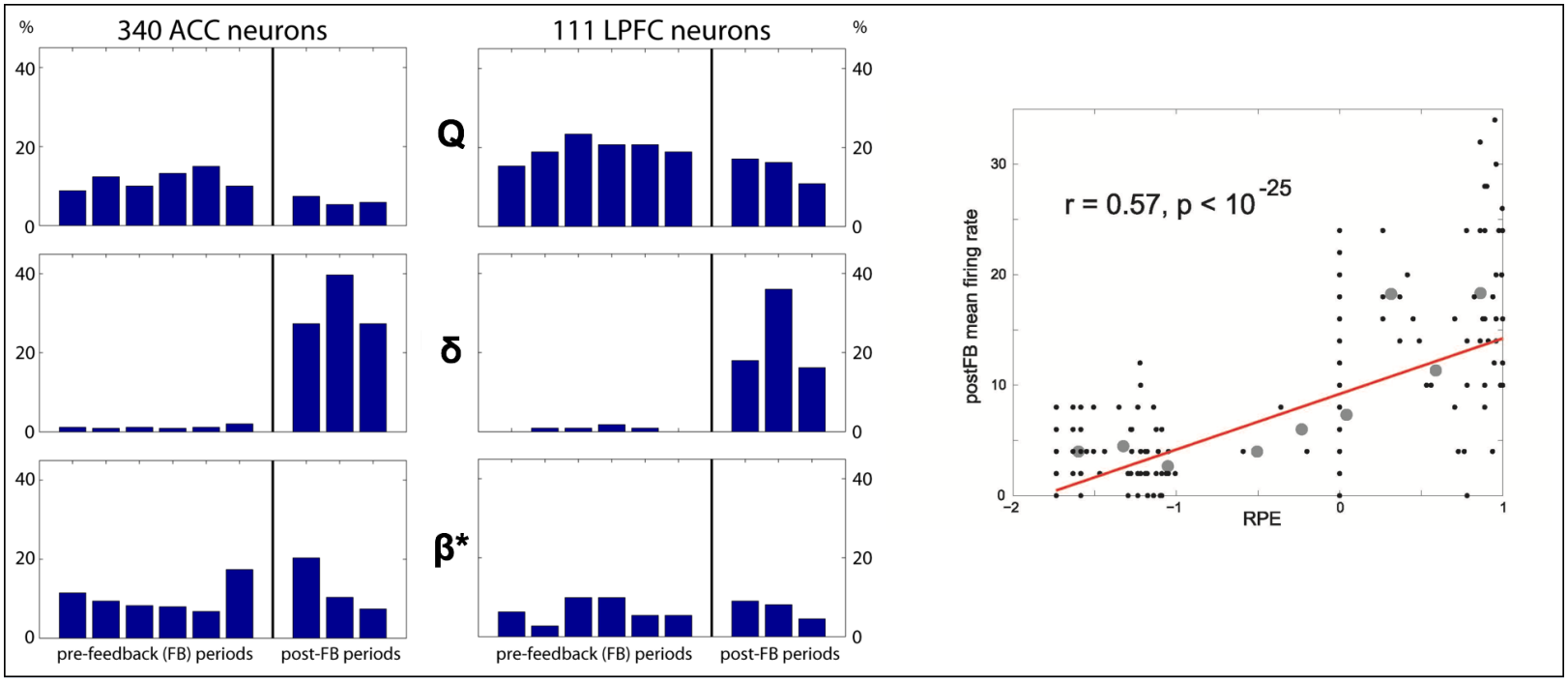
Reproduction of the global properties of monkey behavior in the task



Khamassi et al. (2011) Frontiers in Neurorobotics

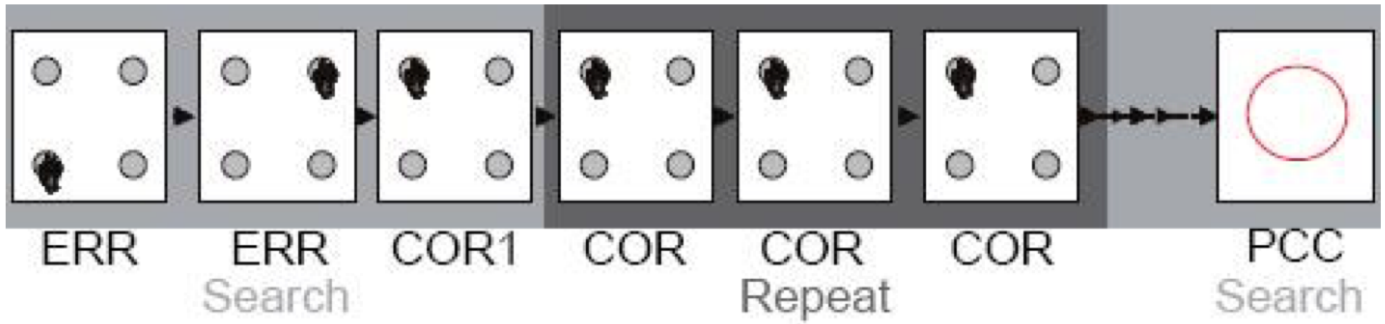
Meta-learning in neuroscience

Correlation between brain activity and model variables



Khamassi et al. (2015) Cerebral Cortex

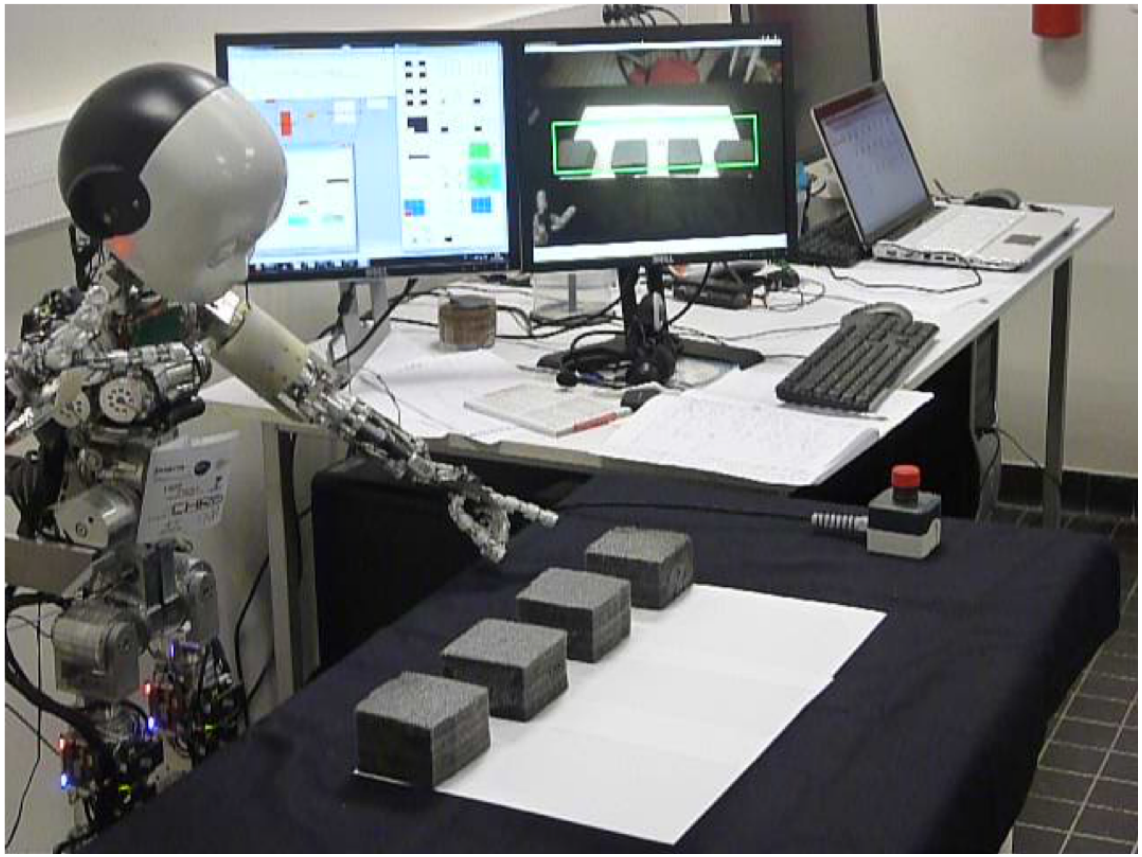
Meta-learning in neuroscience



In the previous task, monkeys and the model a priori 'know' that PCC means a reset of exploration rate and action values.
Here, we want the iCub robot to learn it by itself.

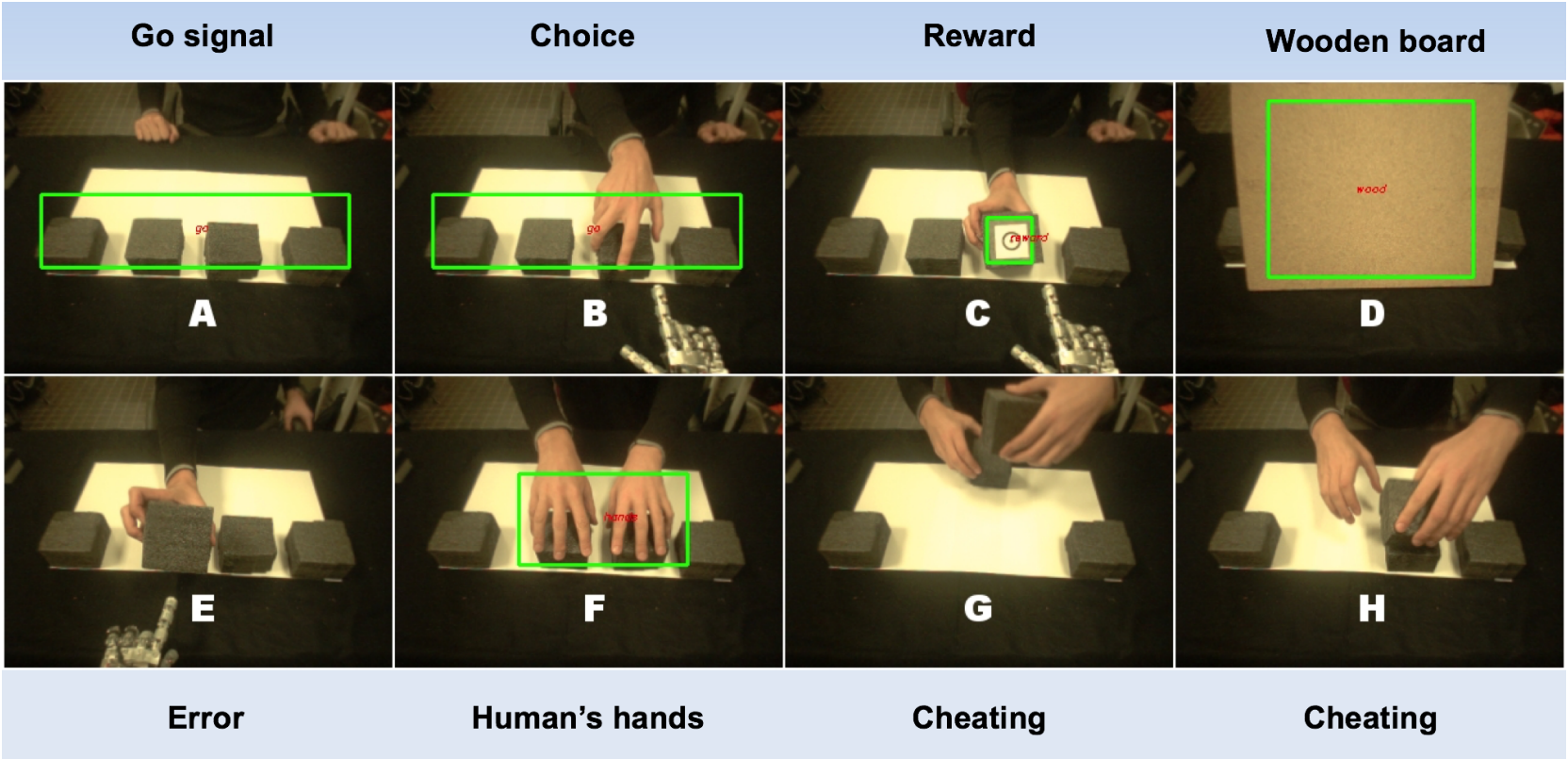
Back to robotics

Meta-learning applied to Human-Robot Interaction



Khamassi et al. (2011) Frontiers in Neurorobotics

Back to robotics

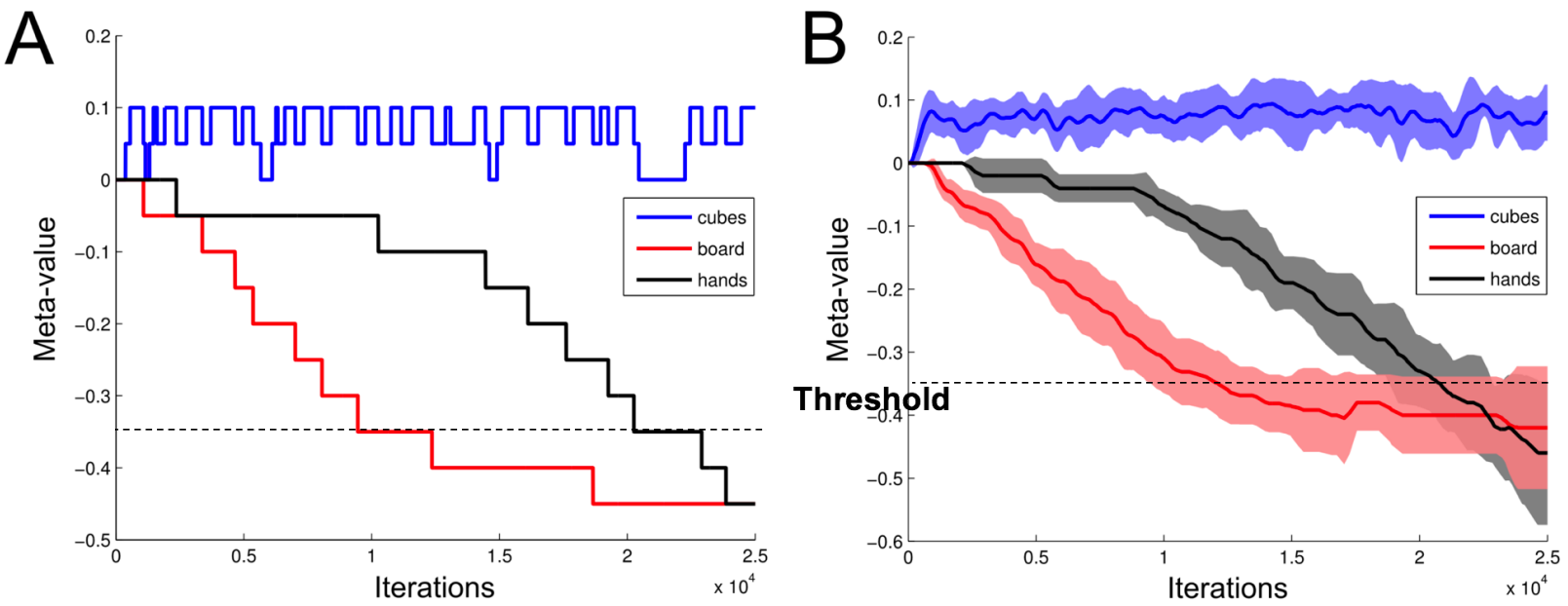


Khamassi et al. (2011) Frontiers in Neurorobotics

Back to robotics

Learning meta-values: $m_i(t+1) = m_i(t) + \alpha' \Delta(\hat{r})$

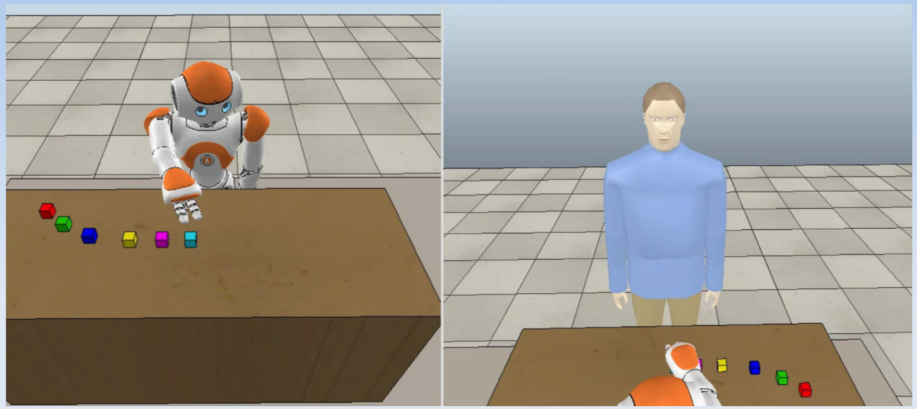
$\hat{r}$: average reward



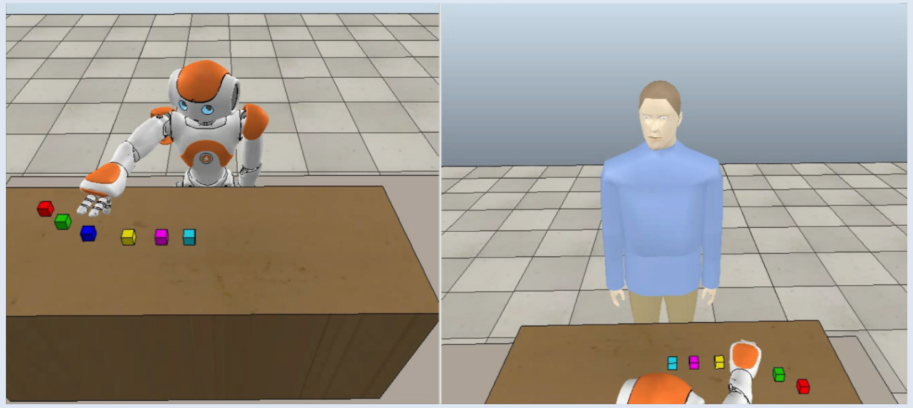
Khamassi et al. (2011) Frontiers in Neurorobotics

Application to child-robot interaction

High human engagement
(in pointing task)



Low human engagement



BABY ROBOT
NEXTGEN SOCIAL ROBOTICS

Khamassi et al. (2018) IEEE Transactions in Cognitive and Developmental Systems

Application to child-robot interaction

Simulated human-robot interaction where the (social) reward function is:

$reward(t) = (1 - \lambda)engagement(t) + \lambda \Delta engagement(t)$ with $\lambda = 0.7$

Estimation of short- and long-term reward running averages with two different time constants τ_1 and τ_2 (Schweighofer & Doya, 2003):

$\Delta \bar{r}(t) = (r(t) - \bar{r}(t))/\tau_1$ and $\Delta \bar{\bar{r}}(t) = (\bar{r}(t) - \bar{\bar{r}}(t))/\tau_2$

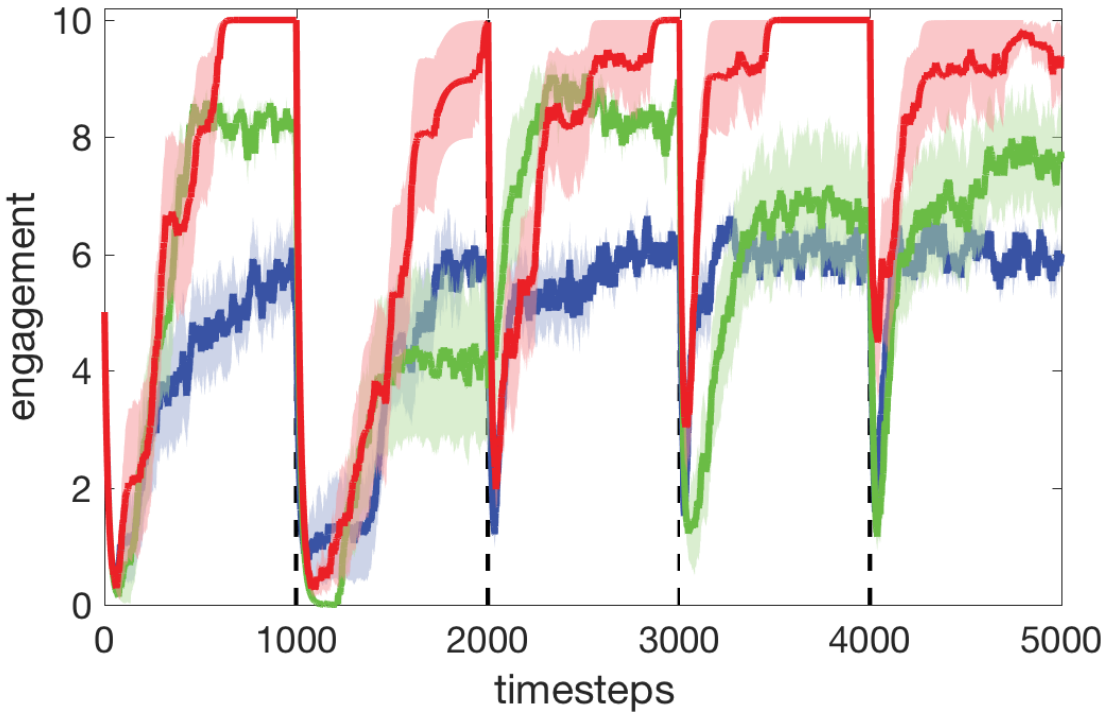
Update of exploration parameters (inverse temperature β_t used in softmax function; and Gaussian width σ_t for continuous action parameters exploration) with:

$\beta_t = F(\mu(\bar{r}(t) - \bar{\bar{r}}(t)))$ and $\sigma_t = G(\mu(\bar{r}(t) - \bar{\bar{r}}(t)))$

where μ is a learning rate, $F(x) > 0$ is affine, $0 < G(x) < 20$ is a sigmoid.

Khamassi et al. (2018) IEEE Transactions in Cognitive and Developmental Systems

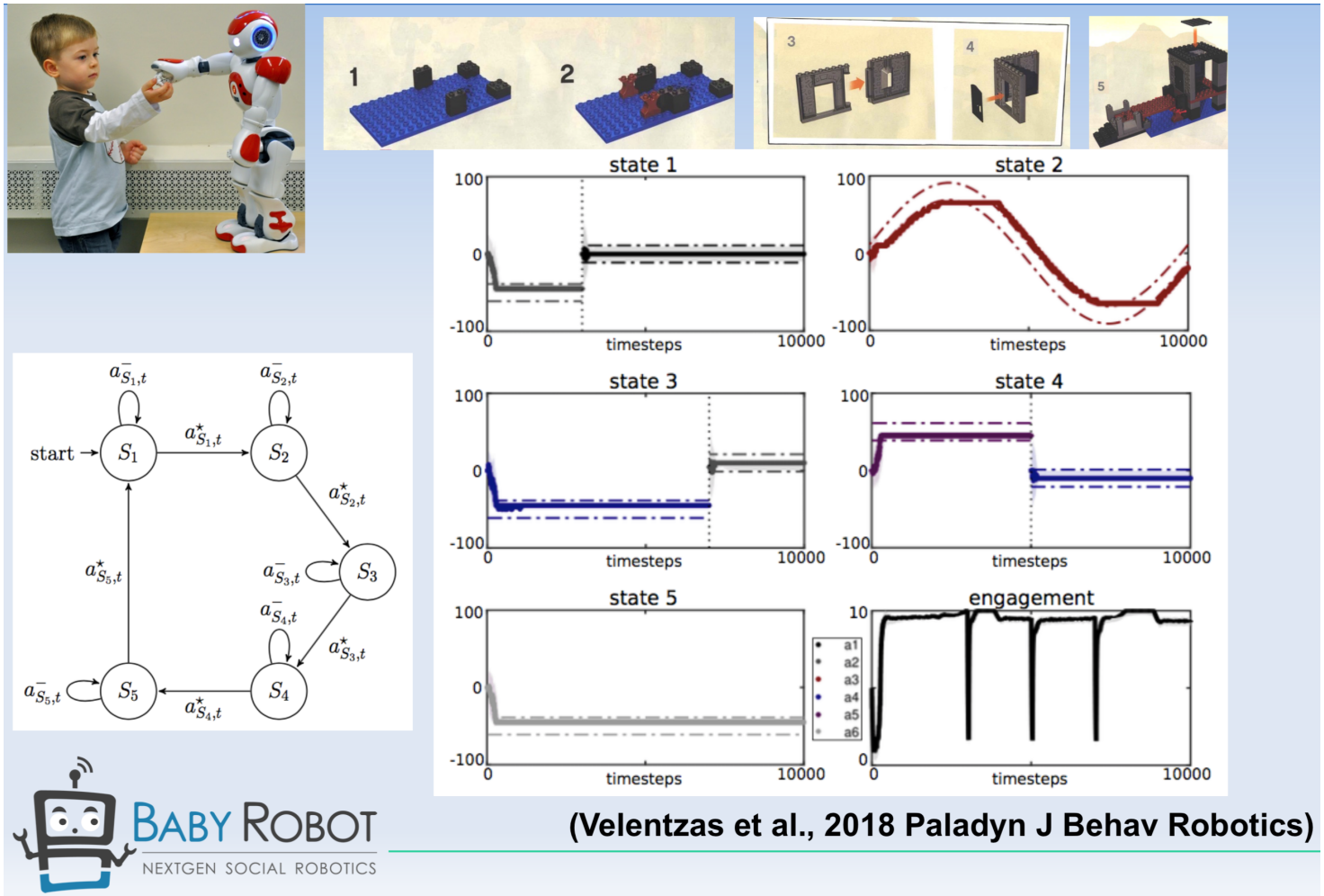
Application to child-robot interaction



Blue: MF-RL; Green: Kalman-TD; Red: Meta-L

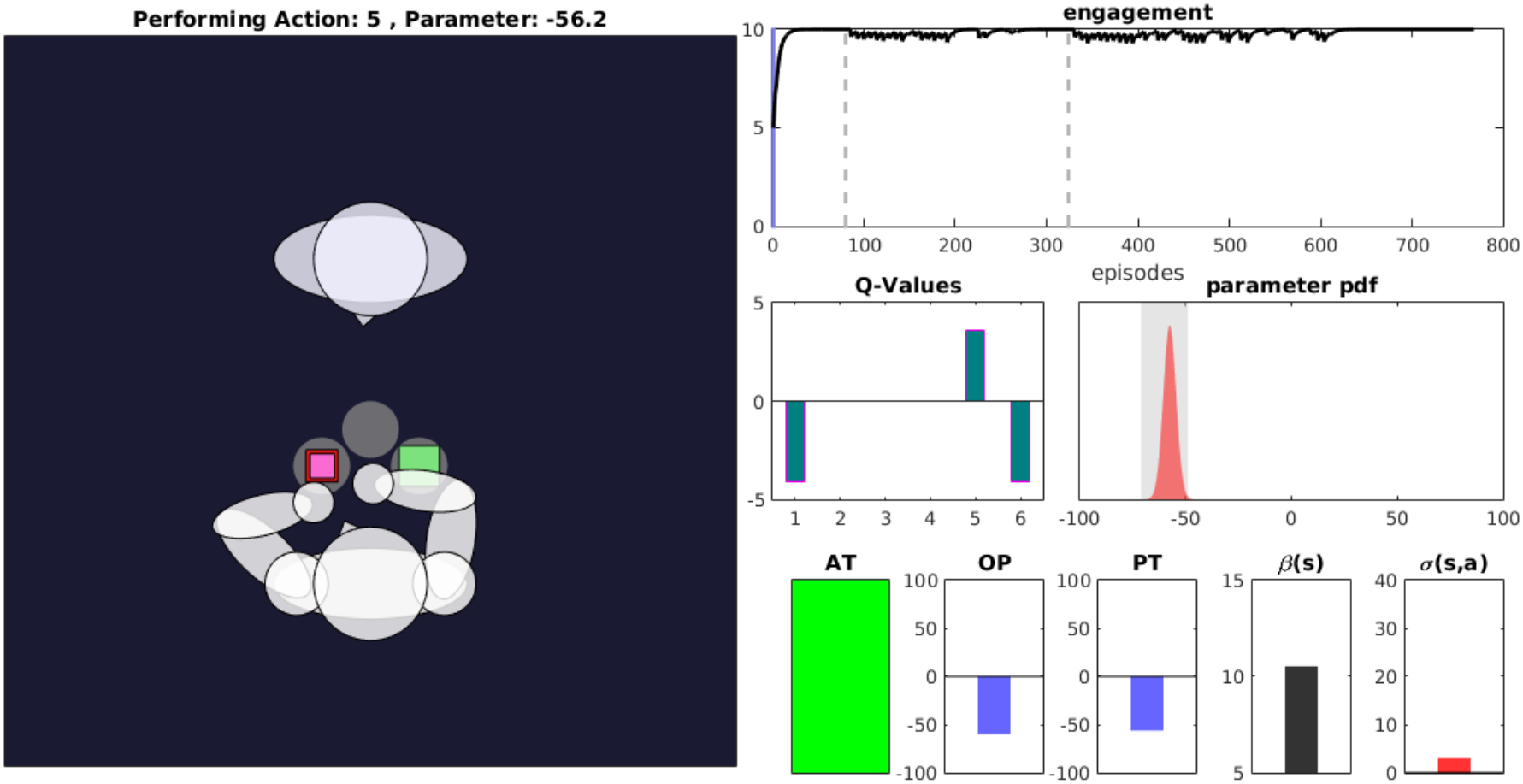
Khamassi et al. (2018) IEEE Transactions in Cognitive and Developmental Systems

Application to child-robot interaction



Velentzas et al. (2018) Paladyn Journal of Behavioral Robotics

Application to child-robot interaction



Velentzas et al. (2018) Paladyn Journal of Behavioral Robotics

Exploration-Exploitation in RL

Classical random exploration in model-free RL models

- Epsilon-greedy: Take the action with the highest value most of the time, and a random action from time to time.
- Softmax: Action probabilities depend on their relative values, normalized with a Boltzmann equation where the inverse temperature β plays the role of a random exploration rate: $P(a_j|s_t) = \frac{exp\big(\beta Q_t(s_t,a_j)\big)}{\sum_a exp(\beta Q_t(s_t,a))}$
- Dynamic random exploration can nevertheless be done with a softmax based on dynamic β_t which is updated as a function of implicit (or explicit) estimations of task volatility or agent performance
[Schweighofer & Doya 2003, Khamassi et al. 2018 IEEE Trans Cog Dev Sys]

Exploration-Exploitation in RL

Directed exploration in bandit and model-free RL models

- Upper Confidence Bound (UCB) methods [Auer et al. 2002] for bandit (single-state) tasks follow the following action selection scheme: choose action a_j which maximizes $Q_t(a_j) + \sqrt{\frac{2 \ln n}{n_{t,j}}}$ where $n_{t,j}$ is the number of times action a_j has been played so far (until timestep t).
- Uncertainty bonuses in Computational neuroscience models are added to Q-value in the softmax: $P(a_j|s_t) = \frac{exp\big(\beta Q_t(s_t,a_j)+\phi\sigma_t(a_j)\big)}{\sum_a exp(\beta Q_t(s_t,a)+\phi\sigma_t(a))}$, where $\sigma_t(a)$ is some measure of uncertainty associated to the estimation of the value of action a, and ϕ is a weighting parameter [Daw, O'Doherty et al. 2006, Frank et al. 2009].
- Experiments show that human do use directed exploration strategies [Wilson et al. 2014, Cogliati-Dezza et al. 2017, Gershman 2018]. But this process might rather be model-based because cognitive load reduces it [Cogliati-Dezza et al. 2019].

CausaL preliminary work

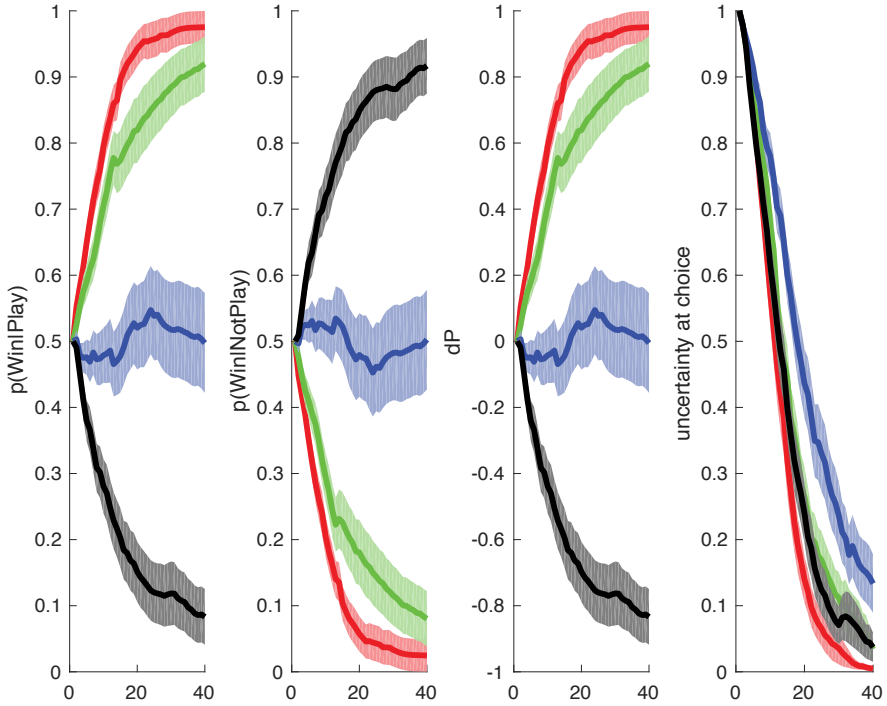
	Model-free RL	Model-based RL
Passive exploration	**e.g., Q-learning model + random exploration** (e.g., epsilon-greedy, softmax, etc.)	**e.g., Real-Time Dynamic Programing (RTDP) + random exploration**
Active (directed) exploration	**e.g., Q-learning model + uncertainty-based exploration bonus**	**### NEW ### RTDP + exploration bonus based on uncertainty in transition model** (contingency uncertainty)

4 simulated models for model-free/model-based RL with passive/active exploration

The most important model (the one in the proposal)

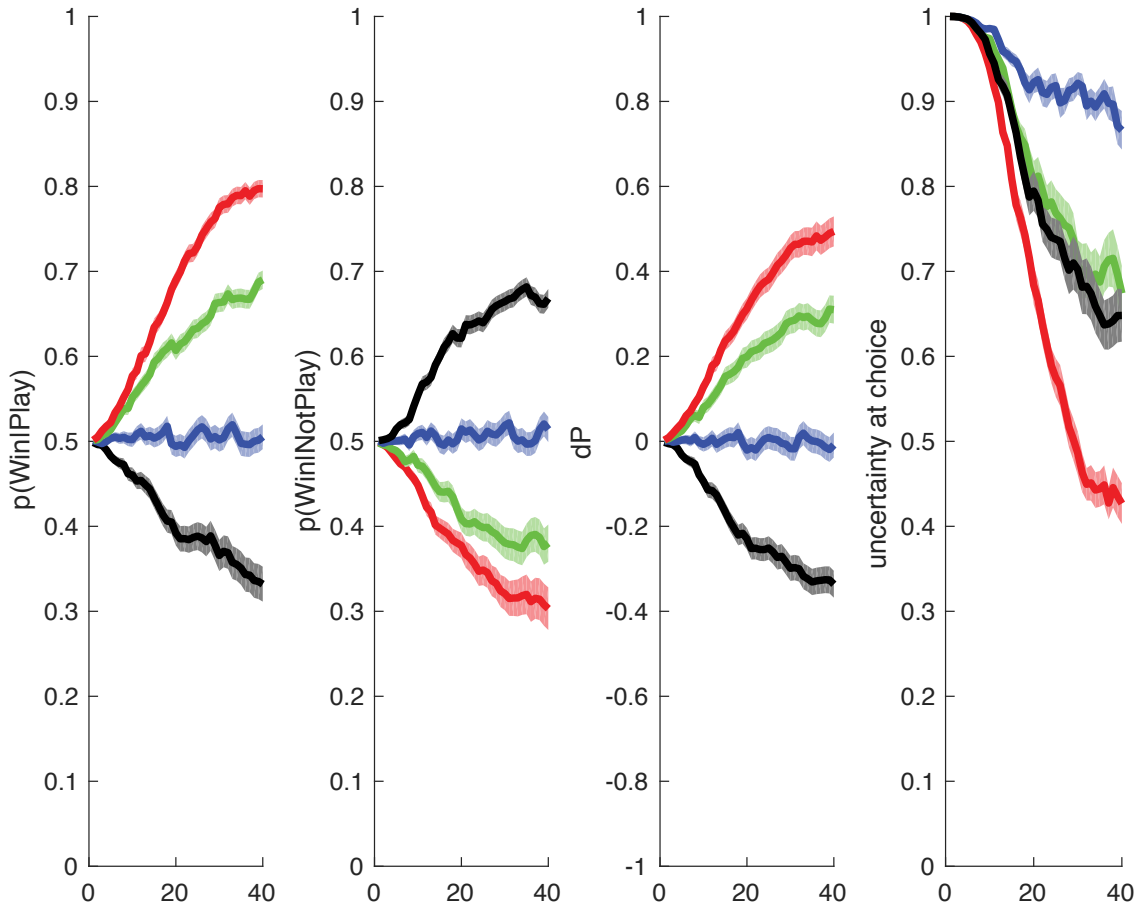
- **model-based RL:** value iteration approximated by a Kalman-filter in order to track both values/probabilities and variances (uncertainties) in the transition and reward functions of the model
- Our measure of uncertainty $\sigma_t(a)$ is the variance in the Kalman-filter associated to the estimate of the reward function $\widehat{\sigma}_{MB,R,t}(s_t, a)$ of the value of action a, and ϕ is a weighting parameter [Daw, O'Doherty et al. 2006, Frank et al. 2009].
- Decision-making based on softmax: $P(a_j|s_t) = \frac{exp\big(\beta Q_{MB,t}(s_t,a_j)+\phi\sigma_t(a_j)\big)}{\sum_a exp\big(\beta Q_{MB,t}(s_t,a)+\phi\sigma_t(a)\big)}$

MFRL+passive exploration simulated in CausaL task

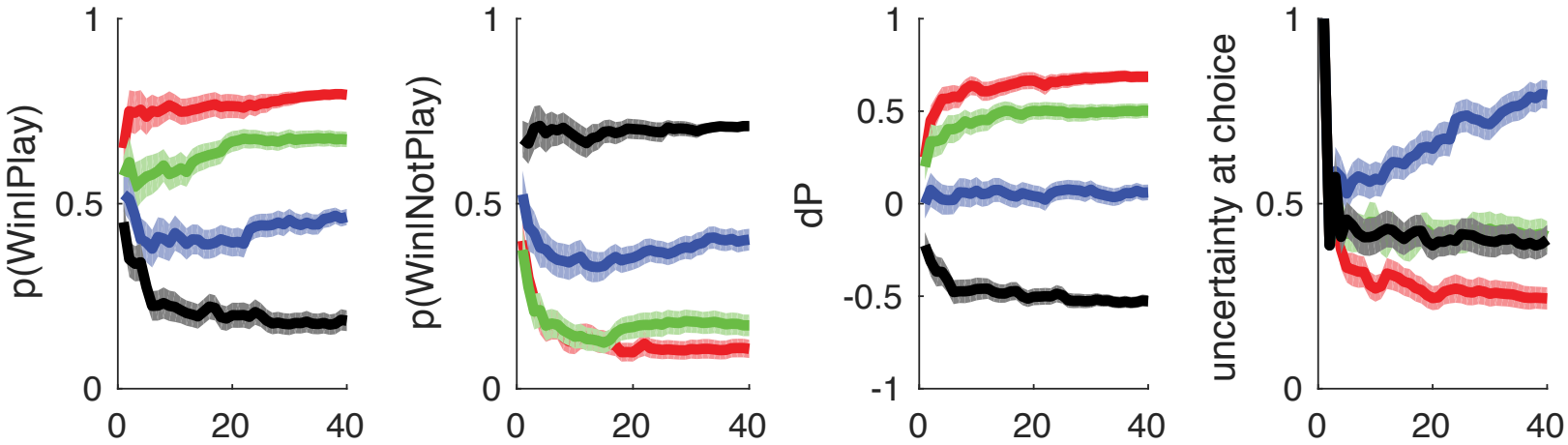


Preliminary simulations of an agent learning the “causal relation” between actions and outcomes with contingency values ΔP = 0.6 (red curve), 0.4 (green), 0 (blue) and -0.4 (black). (A) Evolution of the subjective estimate of $P(Win|Play)$, or $P(O|A)$, as predicted by the model. (B) Subjective estimate of $P(Win|NotPlay)$, or $P(O|NotA)$ (C) Contingency $\Delta P = P(Win|Play) - P(Win|NotPlay)$. (D) Probability of action ’Play’ during the whole learning session.

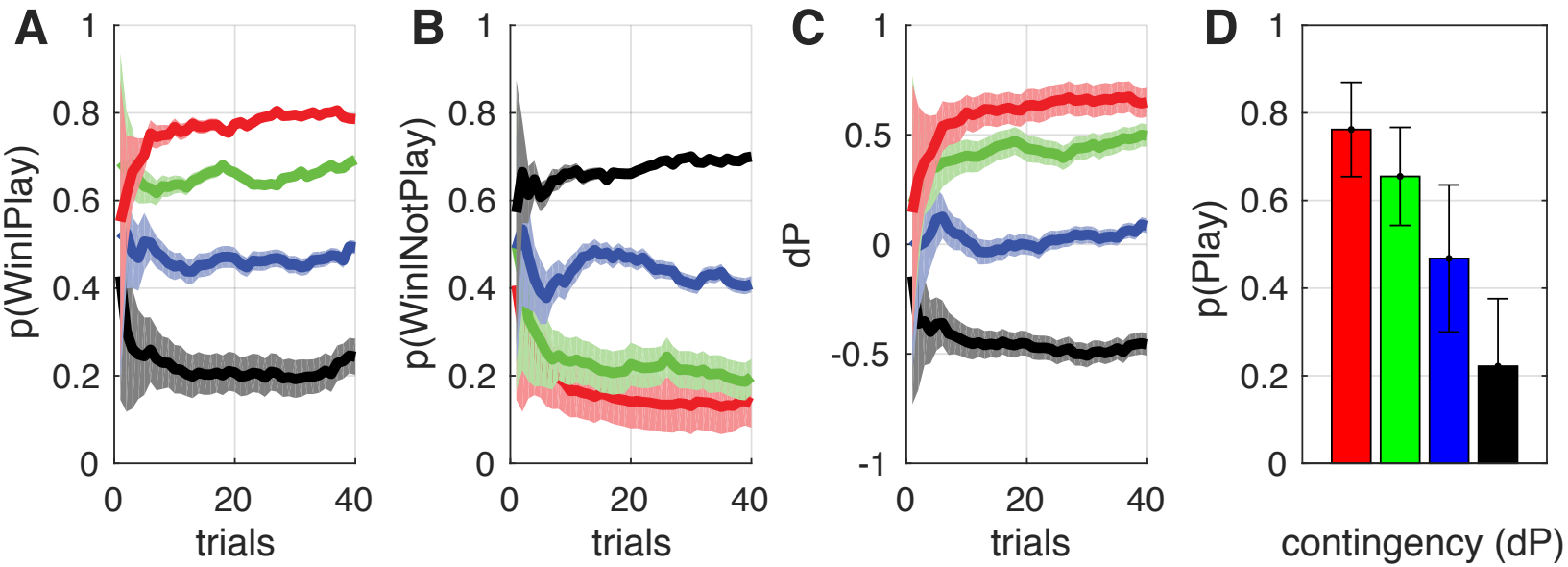
Kalman MFRL+active exploration simulated in CausaL task



MBRL (value iteration)+passive exploration simulated in CausaL task



Kalman MBRL+active exploration simulated in CausaL task



References I

Aubin, L., Khamassi, M., & Girard, B. (2018)

Prioritized Sweeping Neural DynaQ with Multiple Predecessors, and Hippocampal Replays

Living Machines 2018 Conference Paris, France.

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.

Cazé*, R., Khamassi*, M., Aubin, L., & Girard, B. (2018)

Hippocampal replays under the scrutiny of reinforcement learning models

Journal of Neurophysiology To appear.

References II

Coutureau, E., & Killcross, S. (2003)

Inactivation of the infralimbic prefrontal cortex reinstates goal-directed responding in overtrained rats

Behavioural Brain Research 146(1-2), 167–174.

Dollé, L., Chavarriaga, R., Guillot, A., & Khamassi, M. (2018)

Interactions of spatial strategies producing generalization gradient and blocking: A computational approach

PLoS computational biology 14(4), e1006092.

Foster, D. J., & Wilson, M. A. (2006)

Reverse replay of behavioural sequences in hippocampal place cells during the awake state

Nature 440(7084), 680.

Gupta, A. S., van der Meer, M. A., Touretzky, D. S., & Redish, A. D. (2010)

Hippocampal replay is not a simple function of experience

Neuron 65(5), 695-705.

References III

Holroyd, C. B., & McClure, S. M. (2015)

Hierarchical control over effortful behavior by rodent medial frontal cortex: A computational model

Psychological Review 122(1), 54.

Johnson, A., & Redish, A. D. (2007)

Neural ensembles in CA3 transiently encode paths forward of the animal at a decision point

Journal of Neuroscience 27(45), 12176-12189.

Killcross, S., & Coutureau, E. (2003)

Coordination of actions and habits in the medial prefrontal cortex of rats

Cerebral Cortex 13(4), 400–408.

Lee, A. K., & Wilson, M. A. (2002)

Memory of sequential experience in the hippocampus during slow wave sleep

Neuron 36(6), 1183-1194.

References IV

Lin, L.J. (1992)
Self-improving reactive agents based on reinforcement learning, planning and teaching
Machine Learning 8(3-4), 293-321.

Mattar, M., & Daw, N. D. (2018)
Prioritized memory access explains planning and hippocampal replay
Nature Neuroscience X(Y), M-N.

Meyer, J. A., Guillot, A., Girard, B., Khamassi, M., Pirim, P., & Berthoz, A. (2005)
The Psikharpax project: Towards building an artificial rat
Robotics and autonomous systems 50(4), 211-223.

Moore, A. W., & Atkeson, C. G. (1993)
Prioritized sweeping: Reinforcement learning with less data and less time
Machine learning 13(1), 103-130.

Palminteri, S., Lefebvre, G., Kilford, E.J., & Blakemore, S.-J. (2017)
Confirmation bias in human reinforcement learning: Evidence from counterfactual feedback processing

References IV

Peng, J., & Williams, R. J. (1993)

Efficient learning and planning within the Dyna framework

Adaptive Behavior 1(4), 437-454.

Roumis, D. K., & Frank, L. M. (2015)

Hippocampal sharp-wave ripples in waking and sleeping states

Current opinion in neurobiology 35, 6-12.

van Seijen, H., & Sutton, R. S. (2015)

A Deeper Look at Planning as Learning from Replay

Proceedings of the 32nd International Conference on Machine Learning Lille, France.

Sutton, R. S., & Barto, A. G. (1998)

Reinforcement learning: An introduction

MIT press Cambridge, MA.

References V

Doya, K. (2000)

Reinforcement learning in continuous time and space

Neural Computation 12:219-45.

Khamassi, M., Velentzas, G., Tsitsimis, T. & Tzafestas, C. (2018)

Robot fast adaptation to changes in human engagement during simulated dynamic social interaction with active exploration in parameterized reinforcement learning

IEEE Transactions on Cognitive and Developmental Systems 10(4), 881-893.

Keramati, M., & Gutkin, B. (2014)

Homeostatic reinforcement learning for integrating reward collection and physiological stability

eLife 3:e04811.

Konidaris, G., & Barto, A. G. (2006)

Motivational Reinforcement Learning

Springer Simulation of Adaptive Behavior Conference, SAB 2006.

References VI

Schweighofer, N., & Doya, K. (2003)

Meta-learning in Reinforcement Learning

Neural Networks 16:5-9-45.

Auer, P., Cesa-Bianchi, N., & Fischer, P. (2002)

Finite-time Analysis of the Multiarmed Bandit Problem

Machine Learning 47, 235-256.

Daw, N. D., O'doherty, J. P., Dayan, P., Seymour, B., & Dolan, R. J. (2006)

Cortical substrates for exploratory decisions in humans

Nature 441(7095), 876.

Frank, M. J., Doll, B. B., Oas-Terpstra, J., & Moreno, F. (2009)

Prefrontal and striatal dopaminergic genes predict individual differences in exploration and exploitation

Nature Neuroscience 12(8), 1062.

References VII

Cogliati-Dezza, I., Yu, A. J., Cleeremans, A., & Alexander, W. (2017)

Learning the value of information and reward over time when solving exploration-exploitation problems

Scientific reports 7(1), 16919.

Cogliati-Dezza, I., Cleeremans, A., & Alexander, W. (2019)

Should we control? The interplay between cognitive control and information integration in the resolution of the exploration-exploitation dilemma

Journal of Experimental Psychology: General in press.

Wilson, R. C., Geana, A., White, J. M., Ludvig, E. A., & Cohen, J. D. (2014)

Humans use directed and random exploration to solve the explore?exploit dilemma

Journal of Experimental Psychology: General 143(6), 2074.

Gershman, S. J. (2018)

Deconstructing the human algorithms for exploration

Cognition 173, 34-42.

References VIII

Kober, J., Bagnell, J. A., & Peters, J. (2013)

Reinforcement learning in robotics: A survey

The International Journal of Robotics Research 32(11), 1238-1274.

Miller, K. J., Shenhav, A., & Ludvig, E. A. (2019)

Habits without values

Psychological review To appear.

Khamassi, M., & Humphries, M. D. (2012)

Integrating cortico-limbic-basal ganglia architectures for learning model-based and model-free navigation strategies

Frontiers in behavioral neuroscience 6, 79.

Dezfouli, A., & Balleine, B. W. (2012)

Habits, action sequences and reinforcement learning

European Journal of Neuroscience 35(7), 1036-1051.

References VIII

Khamassi, M., Lachèze, L., Girard, B., Berthoz, A., & Guillot, A. (2005)

Actor-Critic models of reinforcement learning in the basal ganglia: from natural to artificial rats

Adaptive Behavior 13(2), 131-148.