Bio-inspired / bio-mimetic action selection & reinforcement learning

Mehdi Khamassi (CNRS, ISIR-UPMC, Paris)

13 September 2016

5AH13 Course, Master Mechatronics for Rehabilitation

University Pierre and Marie Curie (UPMC Paris 6)

REMINDER PREVIOUS COURSES

- Motor control (e.g; how to perform a movement)
- Action selection (e.g. which movement ? which target ?)
- Reinforcement Learning (e.g. some movement lead to « reward » or « punishment »)

→ complementary and interacting processes in the brain.

Important for autonomous and cognitive robots

REMINDER PREVIOUS COURSES

- Motor control (e.g; how to perform a movement)

- Action selection (e.g. which movement ? which target ?)
- Reinforcement Learning (e.g. some movement lead to « reward » or « punishment »)

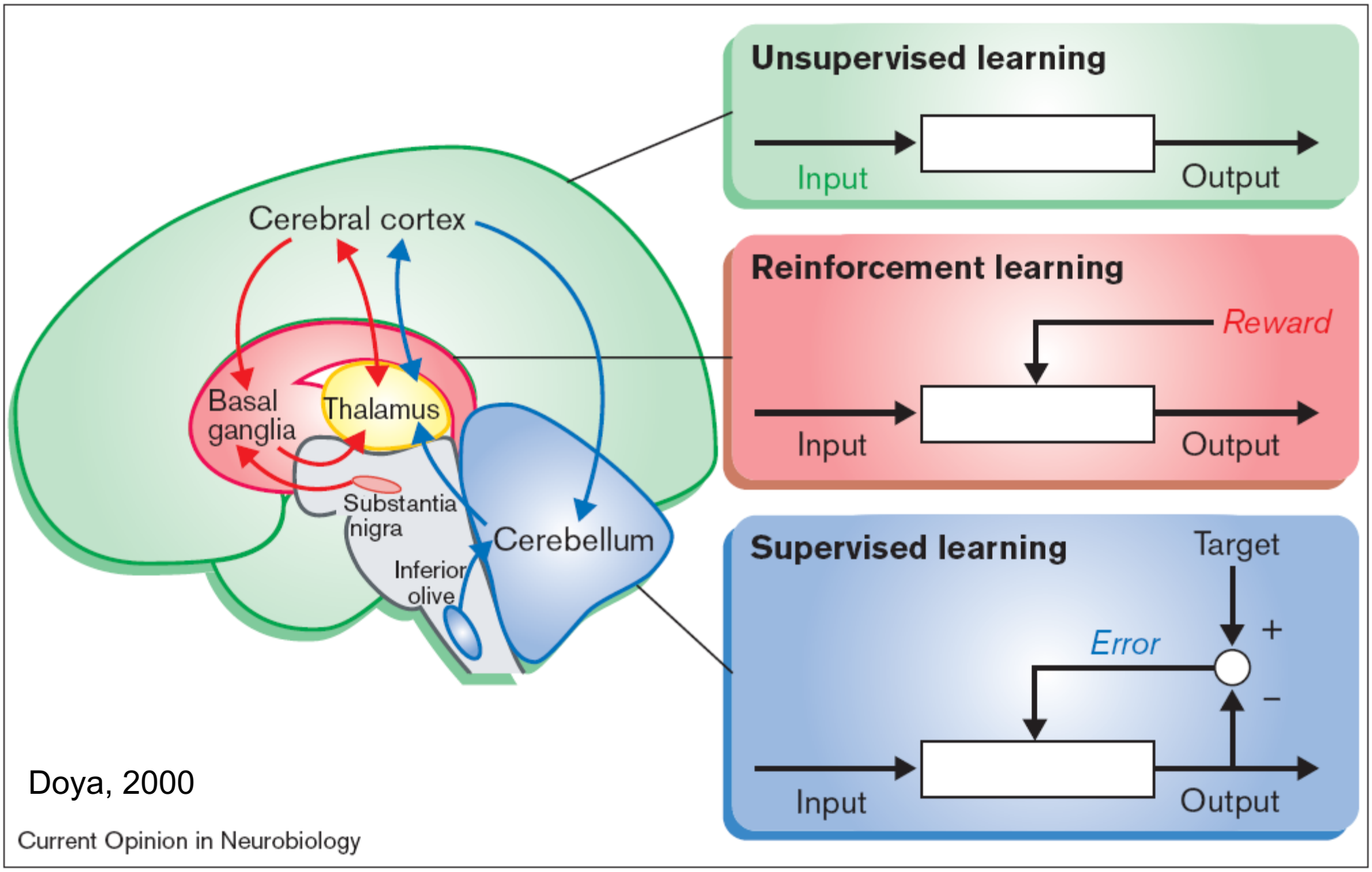
→ complementary and interacting processes in the brain.

Important for autonomous and cognitive robots

OUTLINE

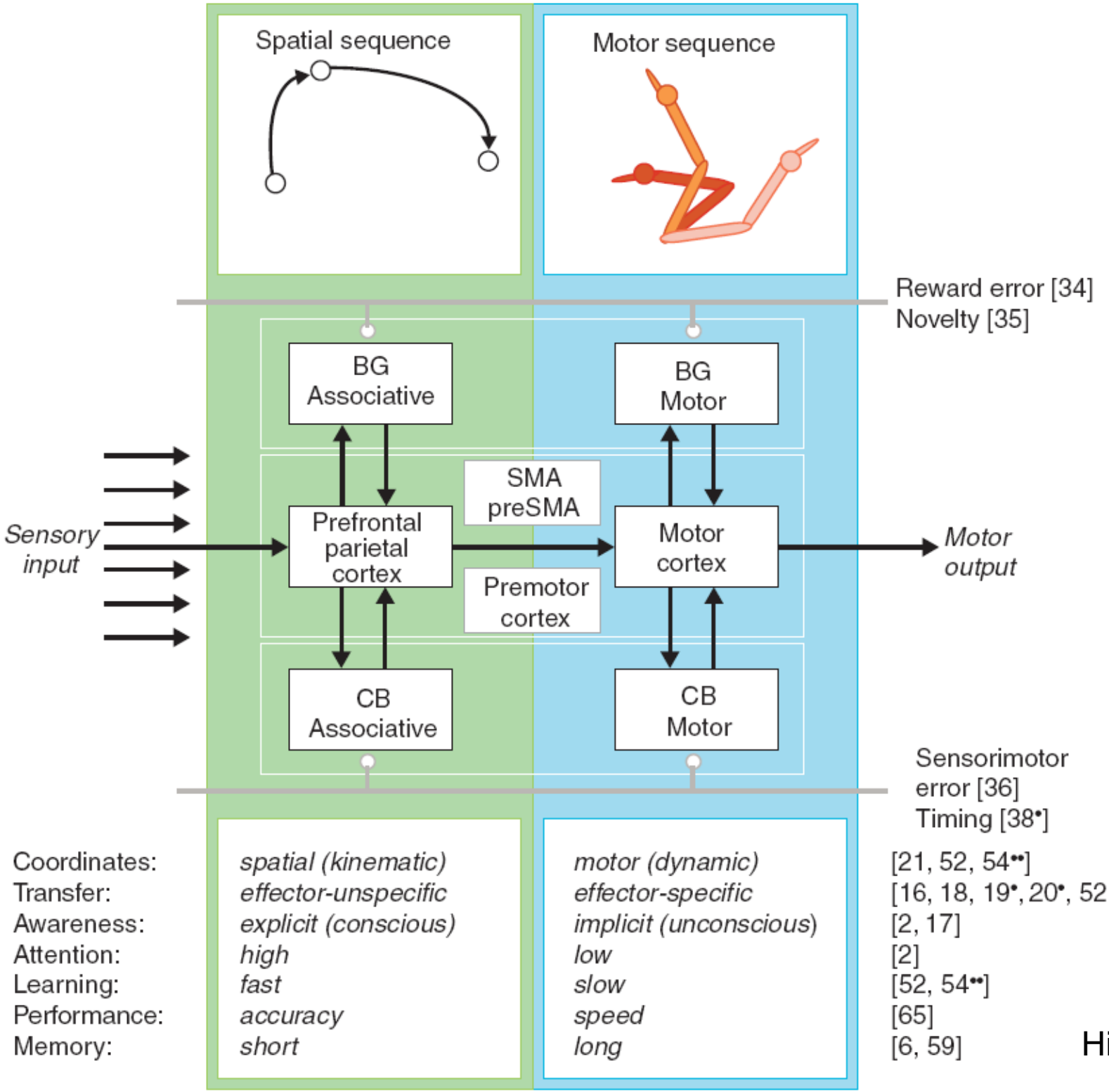
1. Intro
2. Reinforcement Learning model
   - Algorithm
   - Dopamine activity
3. Continuous RL
   - Robot navigation
   - Neuro-inspired models
4. PFC & off-line learning
   - Indirect reinforcement learning
   - Replay during sleep
5. Meta-Learning
   - Principle
   - Neuronal recordings
   - Humanoid Robot interaction

Global organization of the brain



Doya, 2000

Current Opinion in Neurobiology



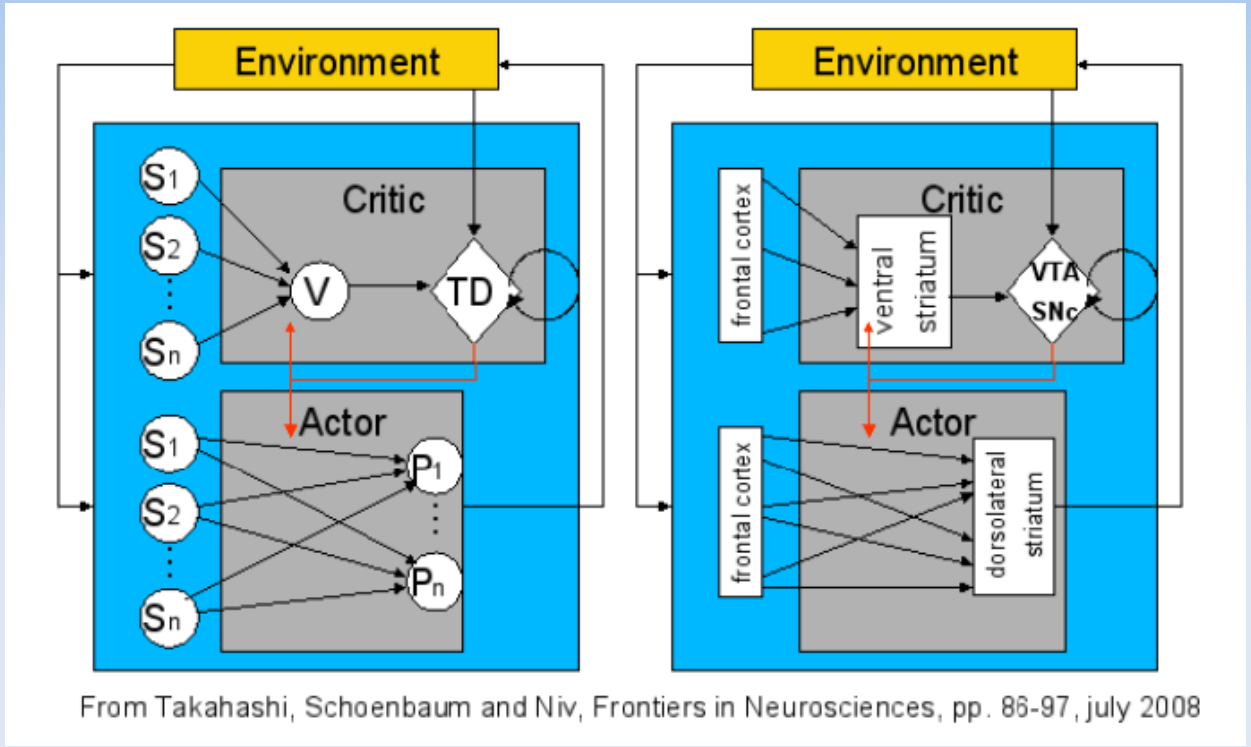
Hikosaka et al., 2002

OUTLINE

1. Intro
2. Reinforcement Learning model
   - Algorithm
   - Dopamine activity
3. Continuous RL
   - Robot navigation
   - Neuro-inspired models
4. PFC & off-line learning
   - Indirect reinforcement learning
   - Replay during sleep
5. Meta-Learning
   - Principle
   - Neuronal recordings
   - Humanoid Robot interaction

THE ACTOR-CRITIC MODEL

Sutton & Barto (1998) Reinforcement Learning: An Introduction



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 constitutes the reinforcement signal.

TD-LEARNING

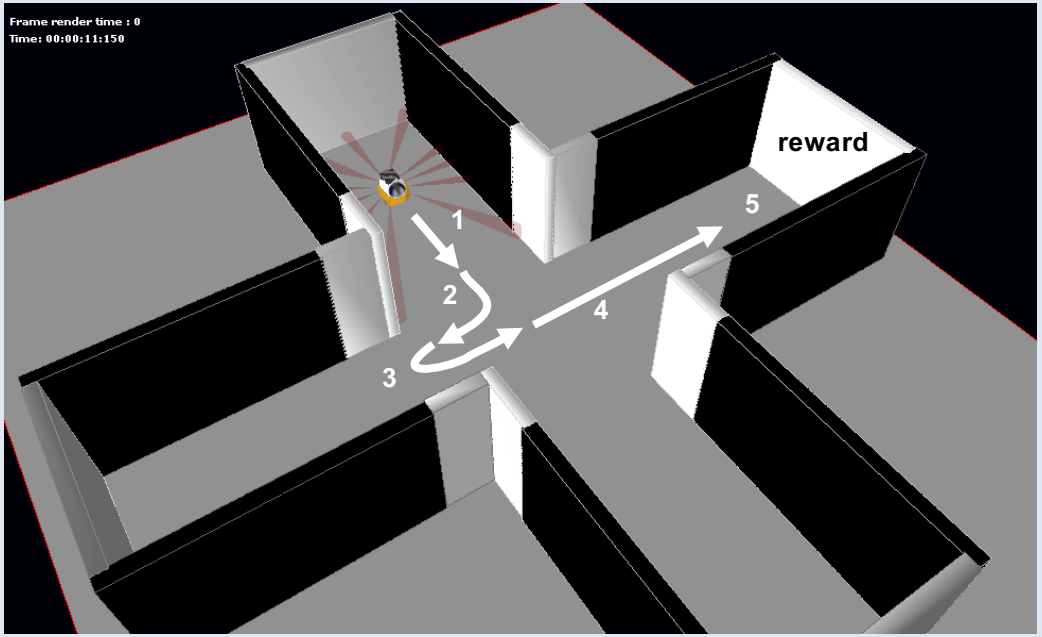
ACTOR	CRITIC	Q-LEARNING
Learns to select actions	Learns to predict reward values	Learns action values

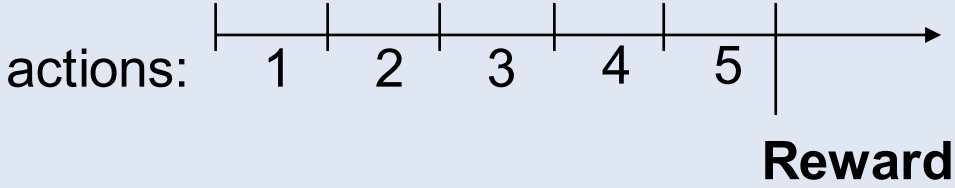
- Developed in the AI community (RL)
- Explains some reward-seeking behaviors (habit learning)
- Resemblance with some part of the brain

  (dopaminergic neurons & striatum)

REINFORCEMENT LEARNING

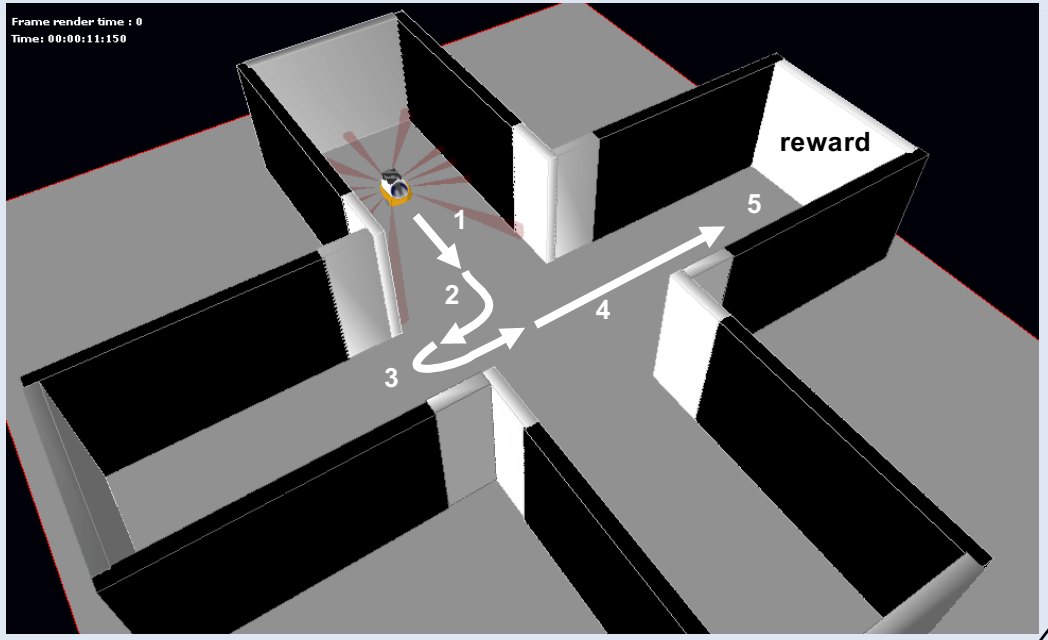
- Learning from delayed reward

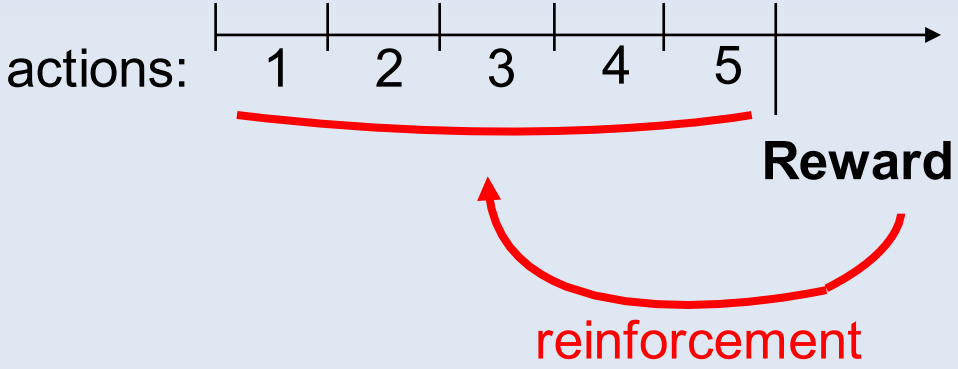




REINFORCEMENT LEARNING

- Learning from delayed reward





$$\delta_t = r_t$$

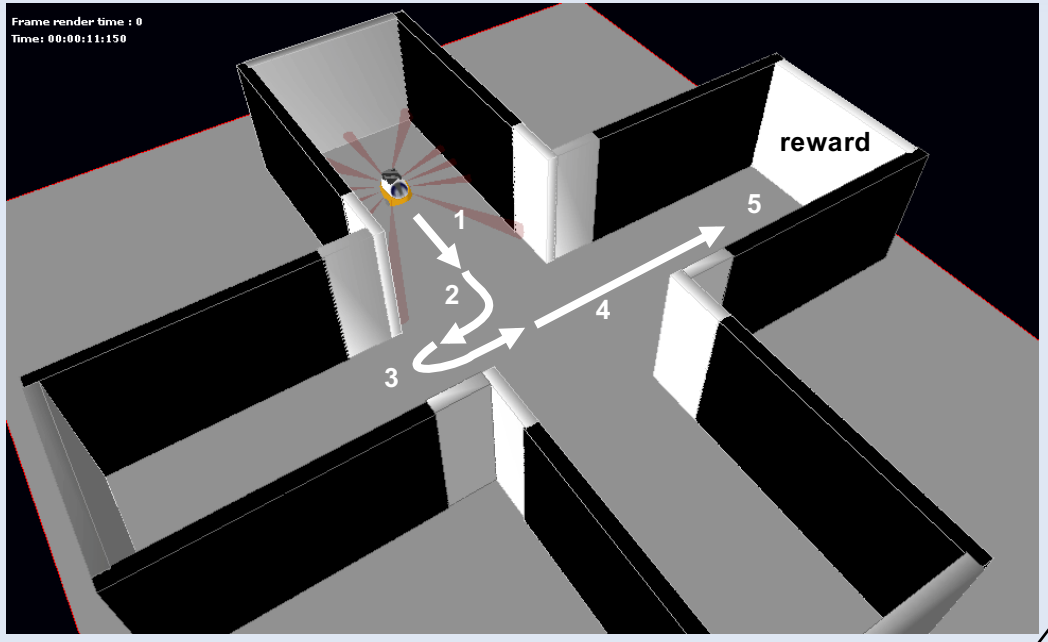
reinforcement

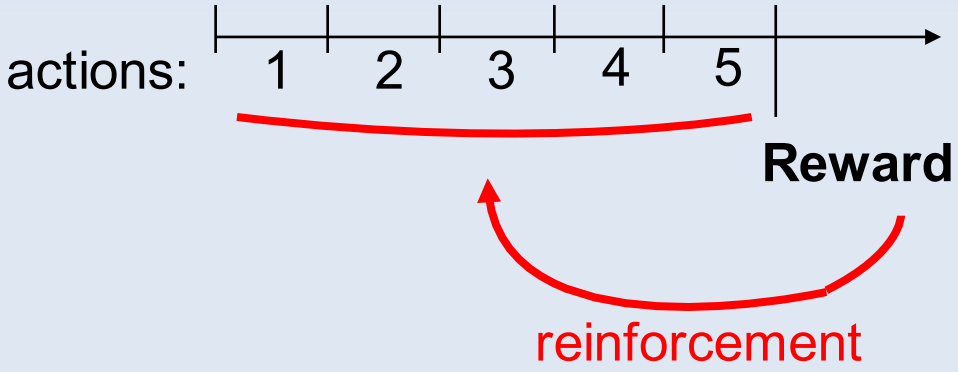
reward

REINFORCEMENT LEARNING

- Learning from delayed reward

Value estimation ("reward prediction"): $V(s_t)$





$$\delta_{t+n} = r_{t+n} - V(s_t)$$

reinforcement

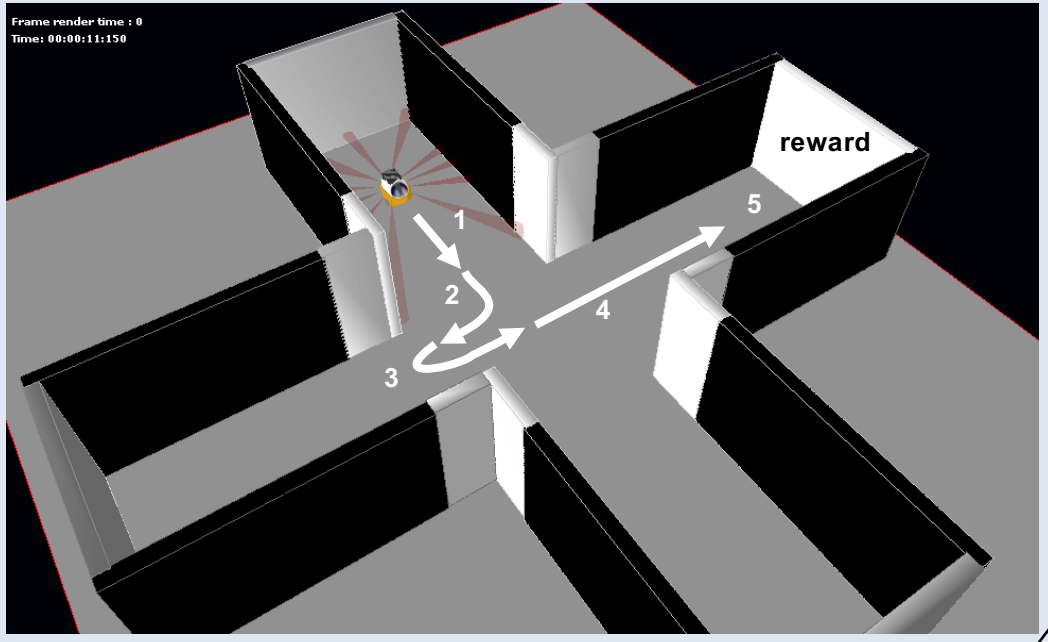
reward

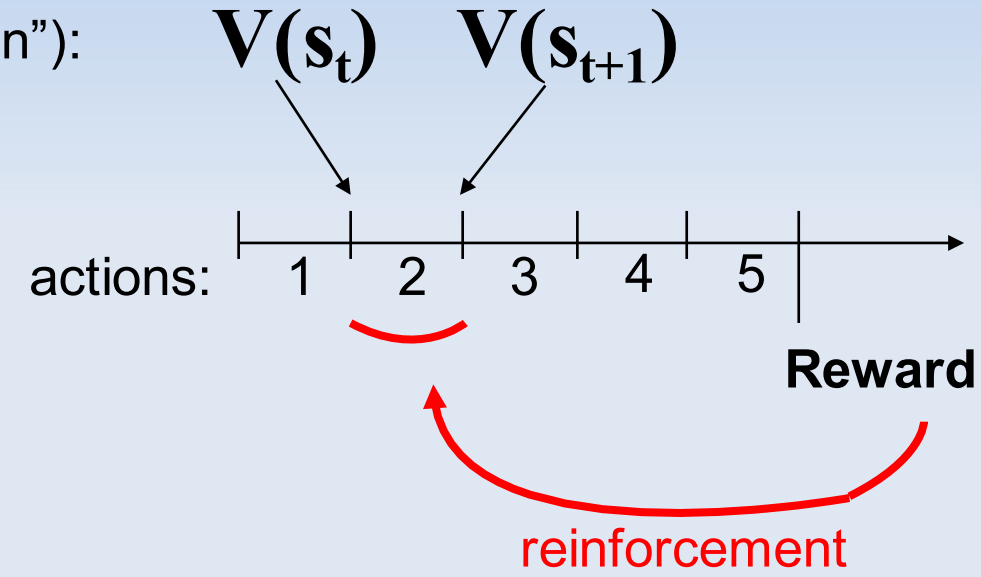
Rescorla and Wagner (1972).

REINFORCEMENT LEARNING

- Temporal-Difference (TD) learning

Value estimation ("reward prediction"):





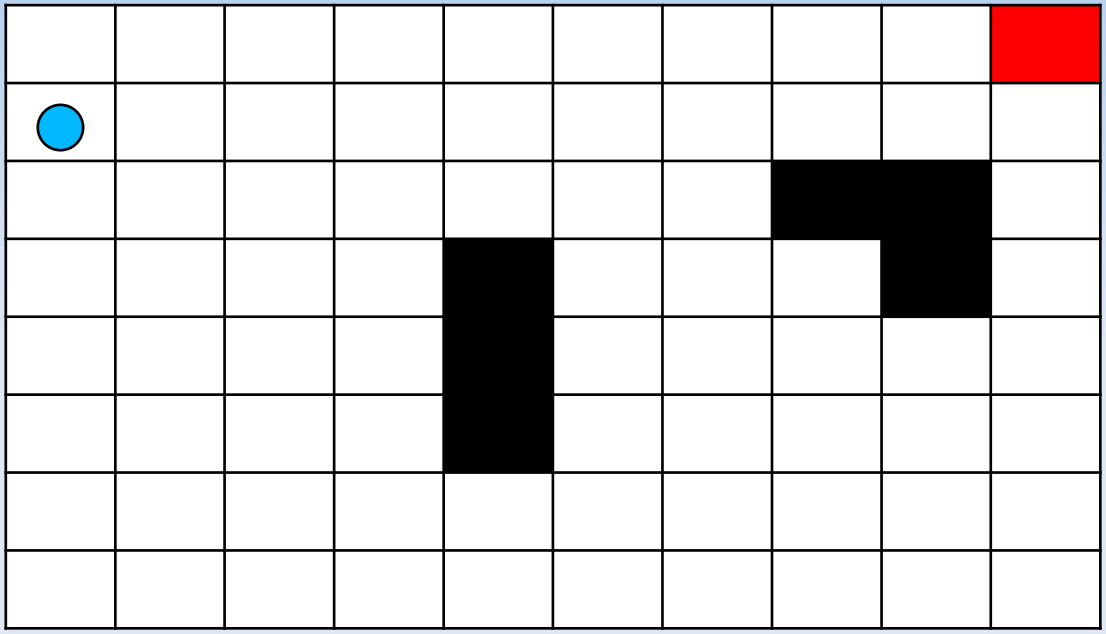
$$\delta_{t+1} = r_{t+1} + \gamma \cdot V(s_{t+1}) - V(s_t) \qquad (\gamma < 1)$$

reinforcement

reward

Sutton and Barto (1998).

REINFORCEMENT LEARNING in a Markov Decision Process



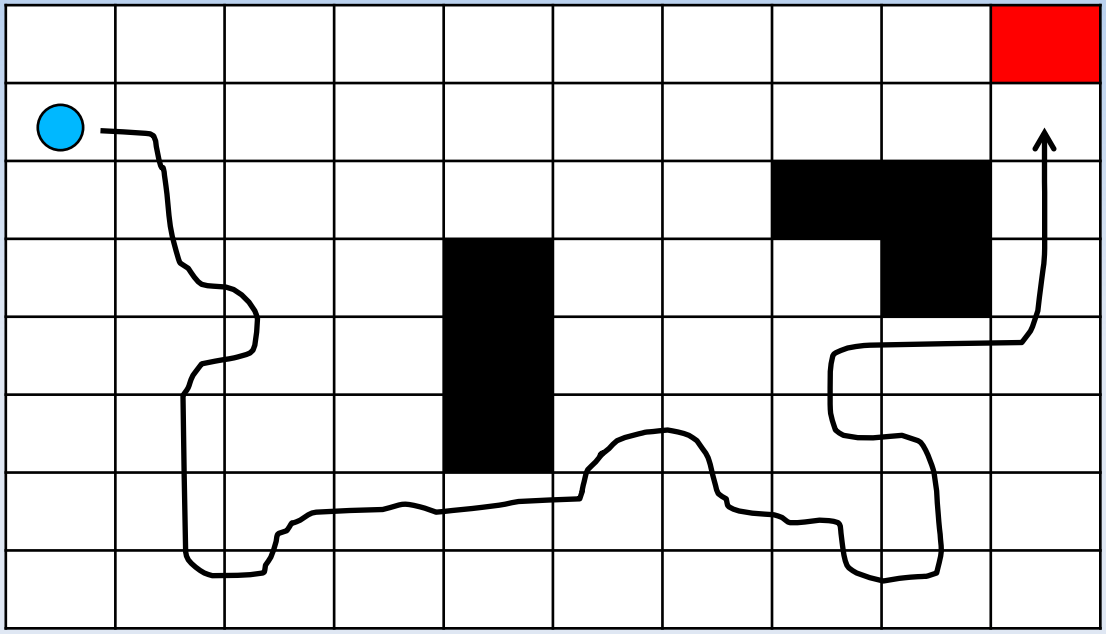
$$\delta_{t+1} = r_{t+1} + \gamma \,.\, V(s_{t+1}) - V(s_t)$$

discount factor (=0.9)

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

learning rate (=0.9)

REINFORCEMENT LEARNING in a Markov Decision Process



0 = 0 + 0 - 0

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

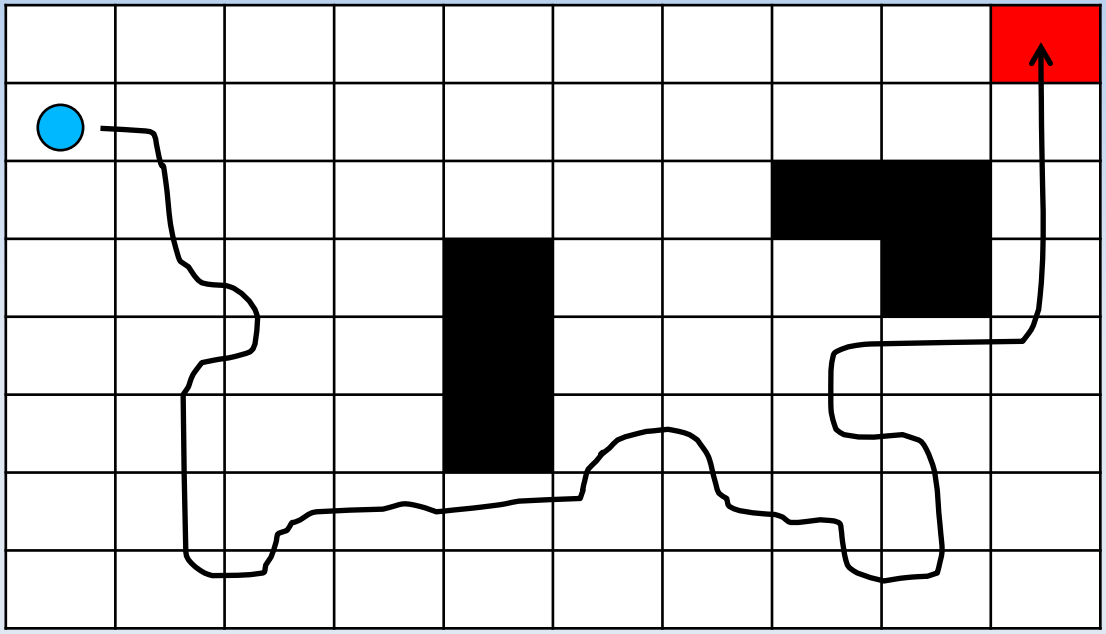
discount factor (=0.9)

0 = 0 + 0.9 * 0

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

learning rate (=0.9)

REINFORCEMENT LEARNING in a Markov Decision Process



1 = 1 + 0 - 0

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

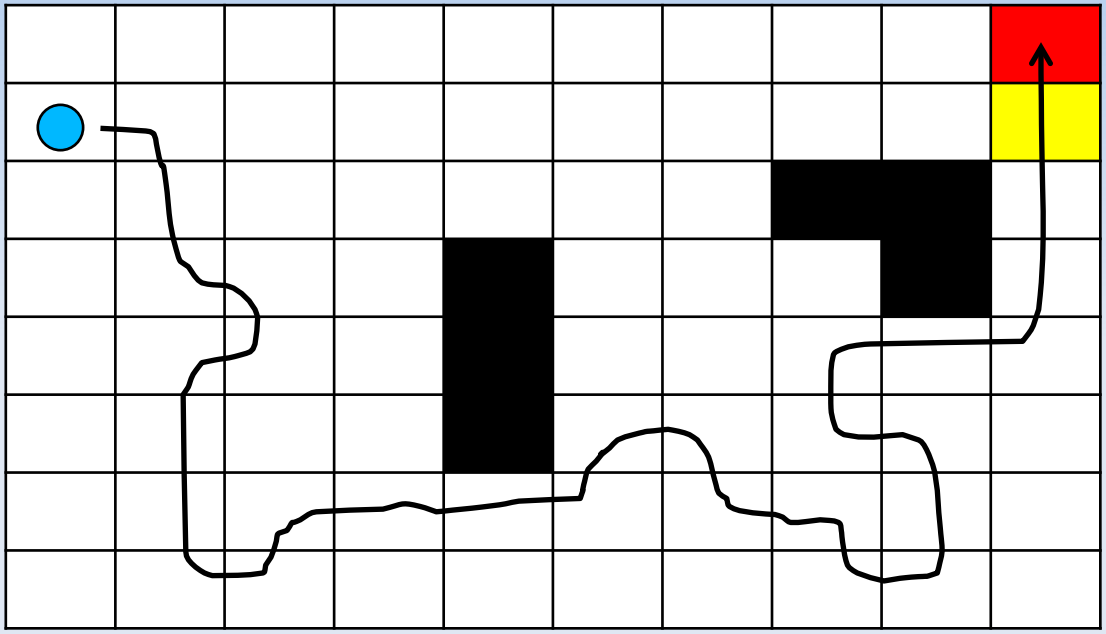
discount factor (=0.9)

0.9 = 0 + 0.9 * 1

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

learning rate (=0.9)

REINFORCEMENT LEARNING in a Markov Decision Process



Color indicates value

1 = 1 + 0 - 0

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

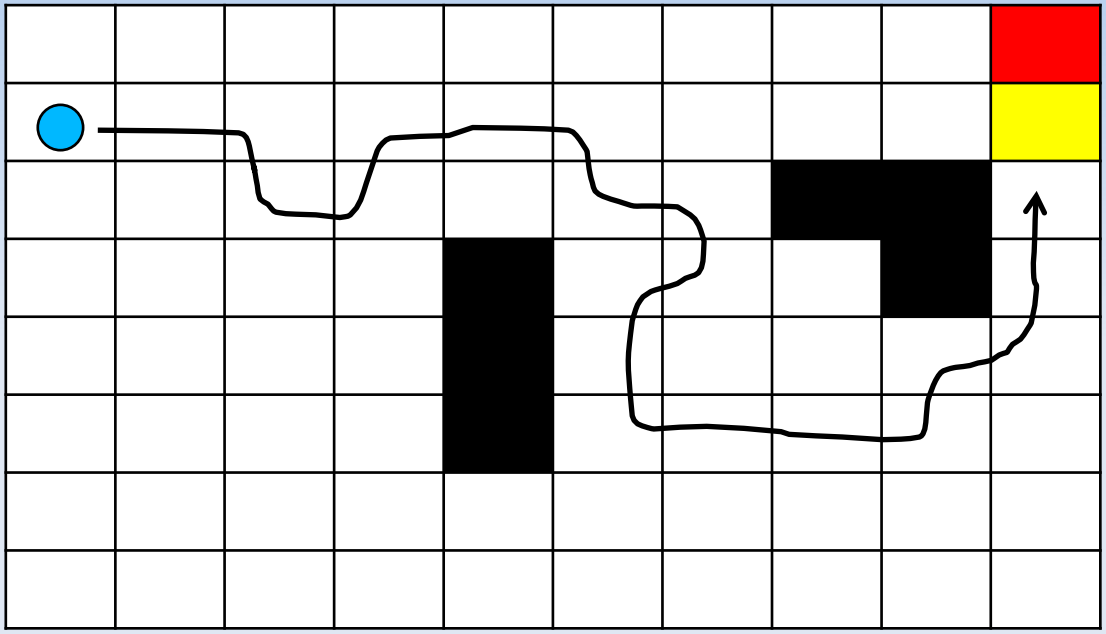
discount factor (=0.9)

0.9 = 0 + 0.9 * 1

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

learning rate (=0.9)

REINFORCEMENT LEARNING in a Markov Decision Process



0 = 0 + 0 - 0

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

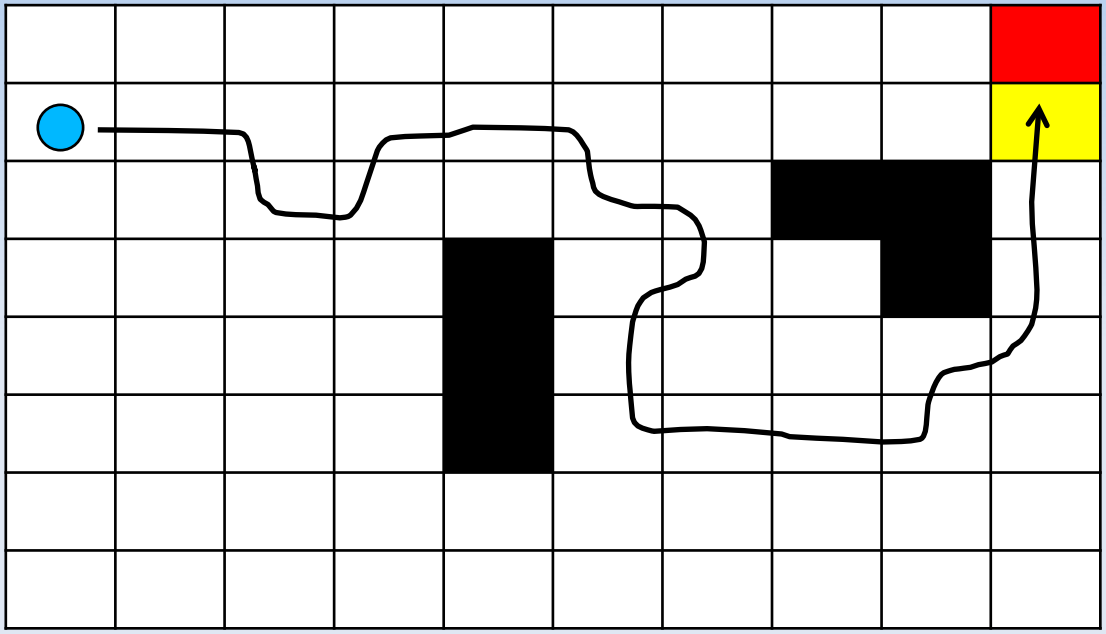
discount factor (=0.9)

0 = 0 + 0.9 * 0

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

learning rate (=0.9)

REINFORCEMENT LEARNING in a Markov Decision Process



0.81 = 0 + 0.9 * 0.9 - 0

0.72 = 0 + 0.9 * 0.81

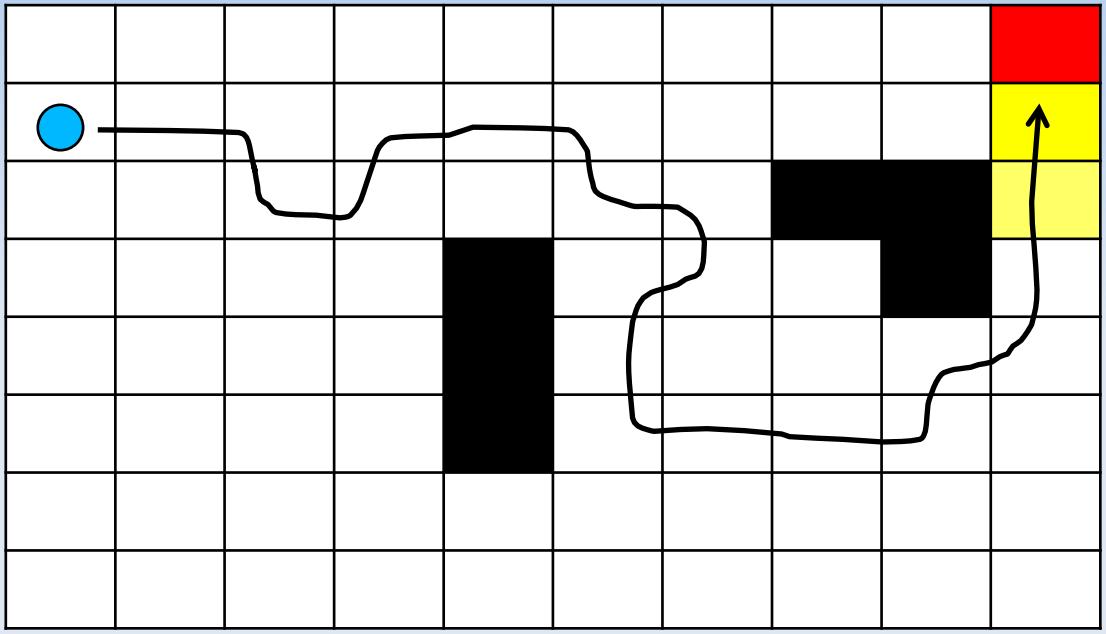
$$\delta_{t+1} = r_{t+1} + \gamma \,.\, V(s_{t+1}) - V(s_t)$$

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

discount factor (=0.9)

learning rate (=0.9)

REINFORCEMENT LEARNING in a Markov Decision Process



Color indicates value

0.81 = 0 + 0.9 * 0.9 - 0

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

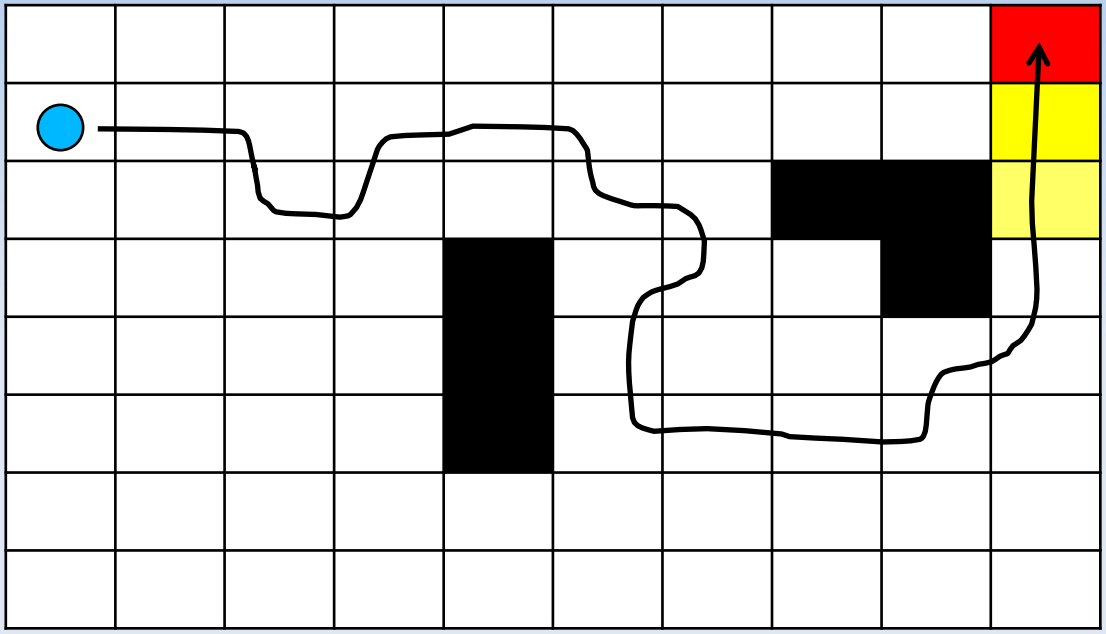
discount factor (=0.9)

0.72 = 0 + 0.9 * 0.81

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

learning rate (=0.9)

REINFORCEMENT LEARNING in a Markov Decision Process



0.1 = 1 + 0 - 0.9

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

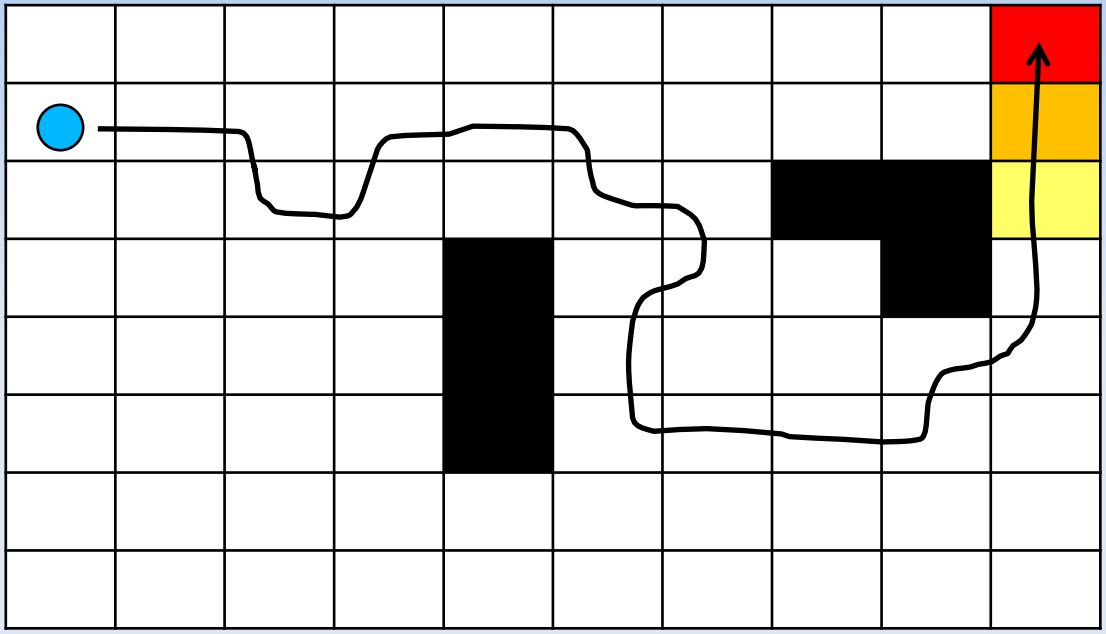
discount factor (=0.9)

0.99 = 0.9 + 0.9 * 0.1

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

learning rate (=0.9)

REINFORCEMENT LEARNING in a Markov Decision Process



Color indicates value

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

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

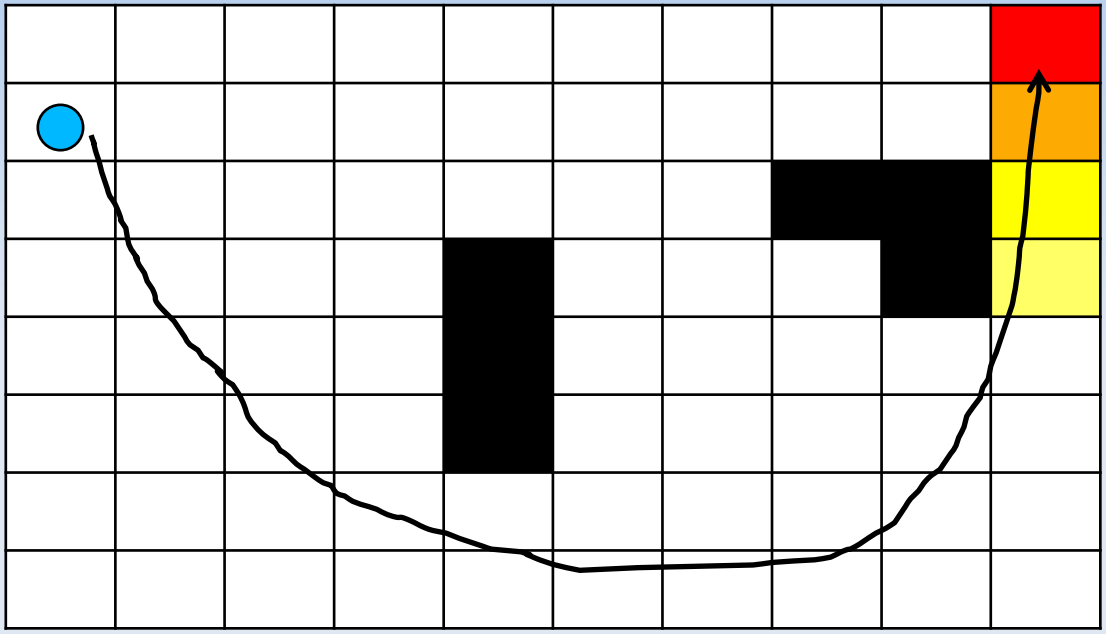
learning rate (=0.9)

REINFORCEMENT LEARNING in a Markov Decision Process

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)

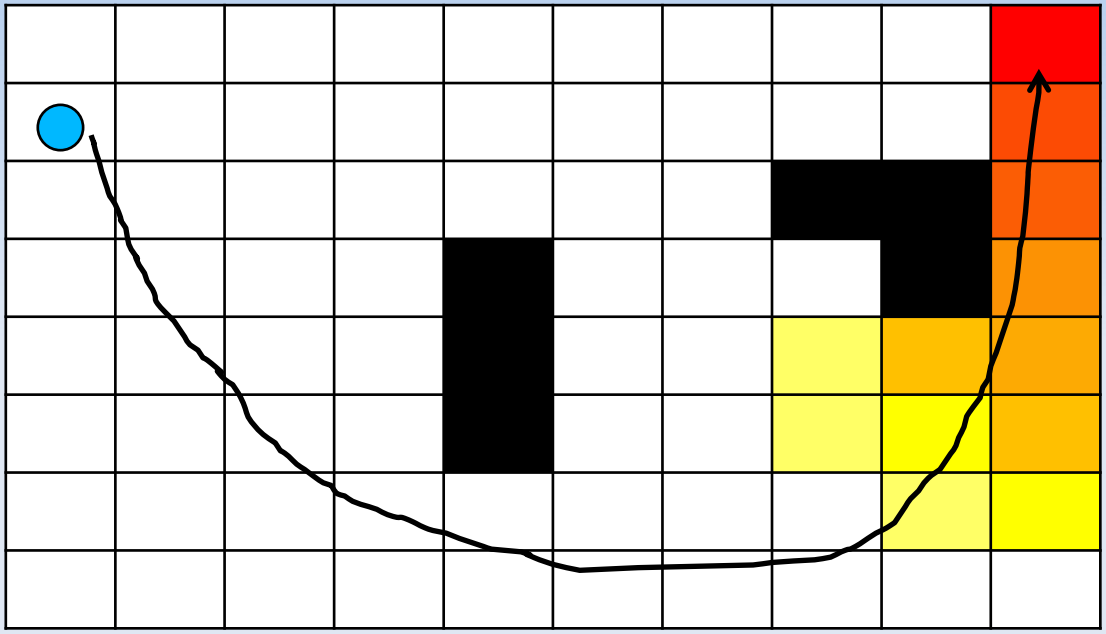
usually small for stability

REINFORCEMENT LEARNING in a Markov Decision Process

After

N simulations

Very long!



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

discount factor (=0.9)

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

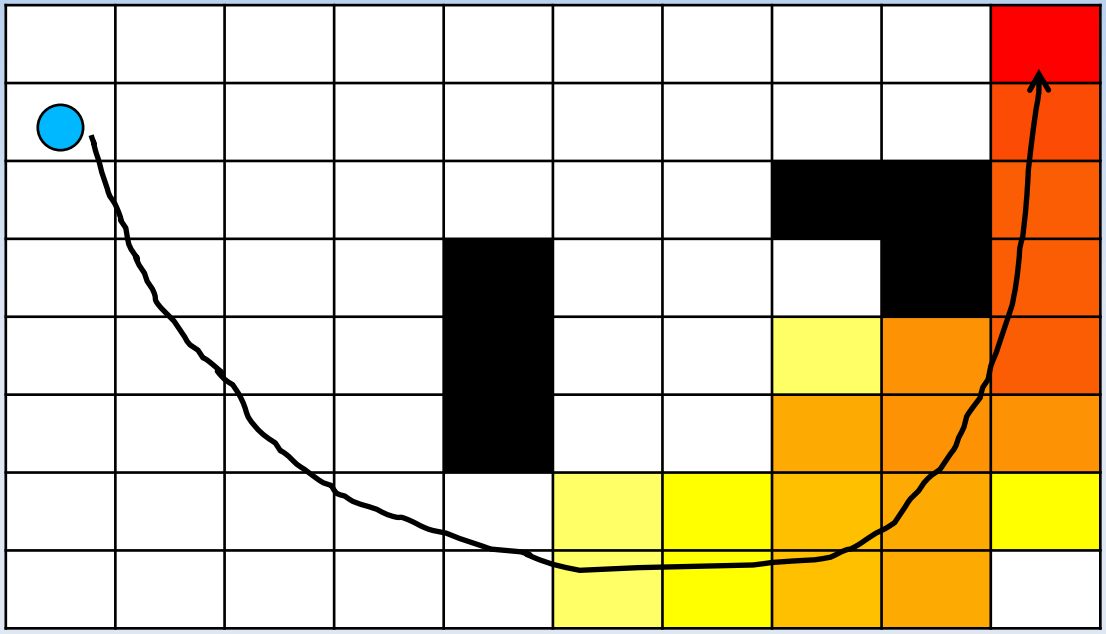
learning rate (=0.1)

REINFORCEMENT LEARNING in a Markov Decision Process

After

N simulations

Very long!



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

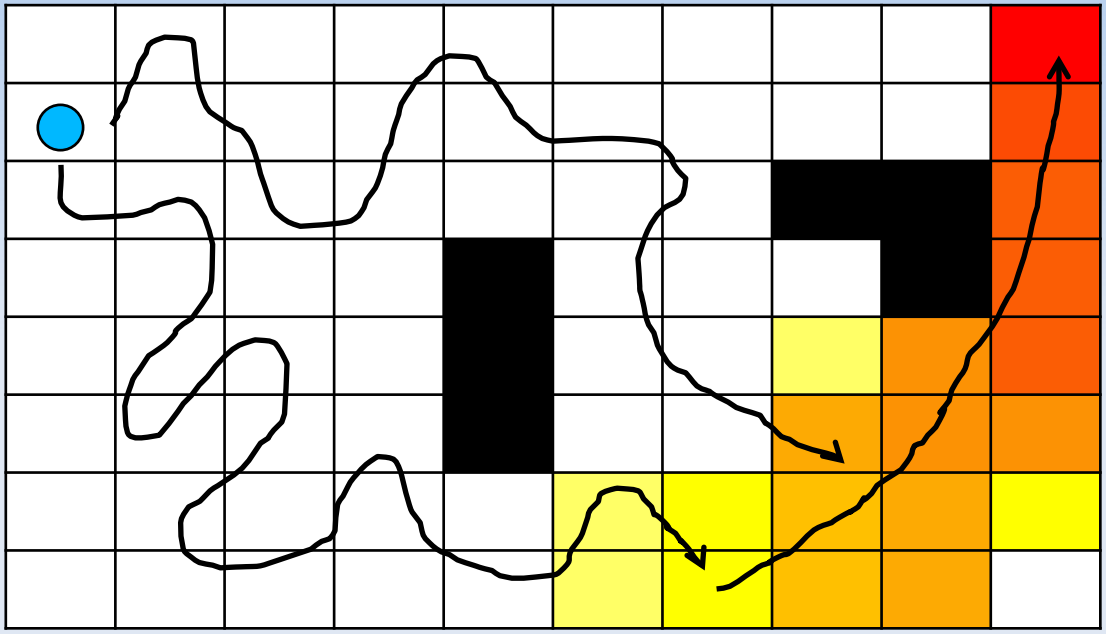
discount factor (=0.9)

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

learning rate (=0.1)

REINFORCEMENT LEARNING in a Markov Decision Process

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)

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

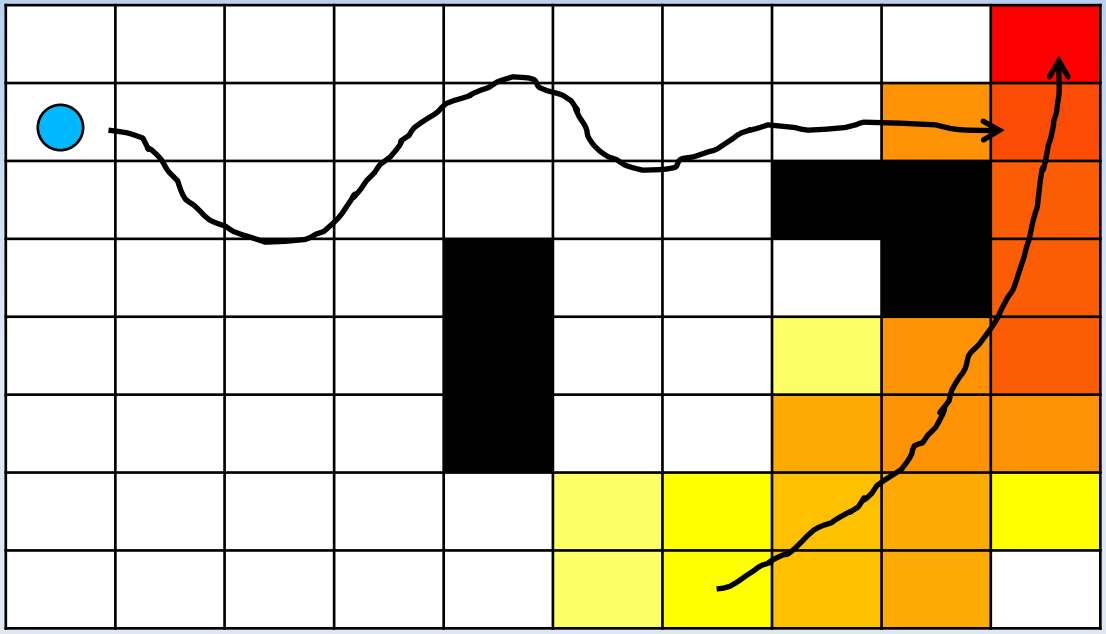
learning rate (=0.1)

REINFORCEMENT LEARNING in a Markov Decision Process

Exploration-

Exploitation

trade-off



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

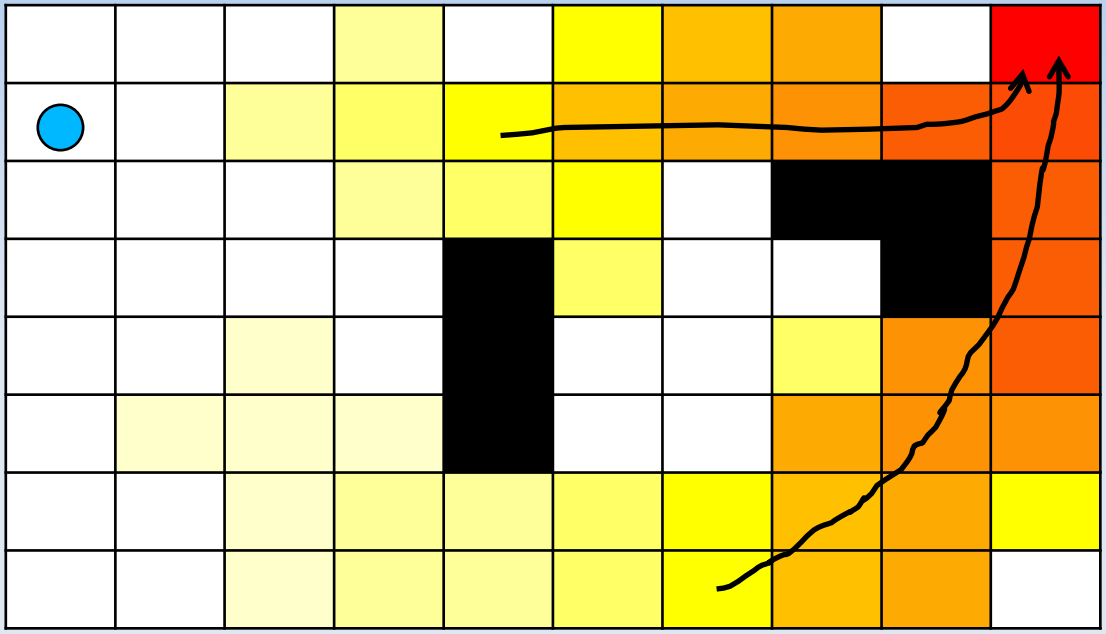
discount factor (=0.9)

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

learning rate (=0.1)

REINFORCEMENT LEARNING in a Markov Decision Process

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)

How can the agent learn a policy?

How to learn to perform the right actions

How can the agent learn a policy?

How to learn to perform the right actions

$\boldsymbol{S}$: state space

$\boldsymbol{A}$: action space

Policy function $\boldsymbol{\pi} : S \longrightarrow A$

What we learned until now:

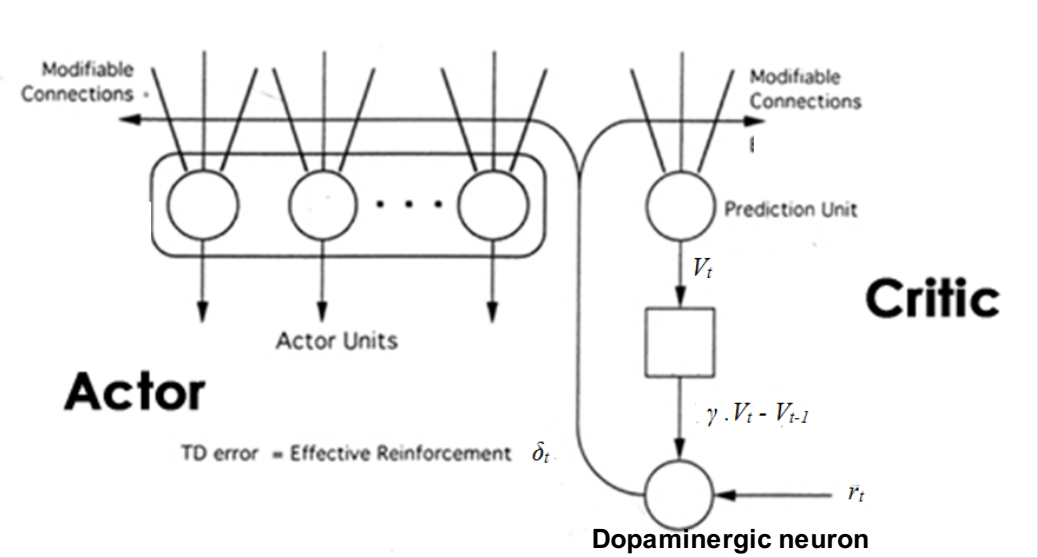
Value function $\boldsymbol{V} : S \longrightarrow R$

The Actor-Critic model

How can the agent learn a policy?

How to learn to perform the right actions

a solution: parallely update a policy and a value function



$$P^{\pi}(a_t|s_t) = P^{\pi}(a_t|s_t) + \alpha \, . \, \delta_{t+1}$$

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

The Q-learning model

How can the agent learn a policy?

How to learn to perform the right actions

other solution: learning Q-values (qualities)

$Q : (S,A) \longrightarrow R$

Q-table:

state / action	a1 : North	a2 : South	a3 : East	a4 : West
s1	0.92	0.10	0.35	0.05
s2	0.25	0.52	0.43	0.37
s3	0.78	0.9	1.0	0.81
s4	0.0	1.0	0.9	0.9
...	...	...	...	...

The Q-learning model

How can the agent learn a policy?

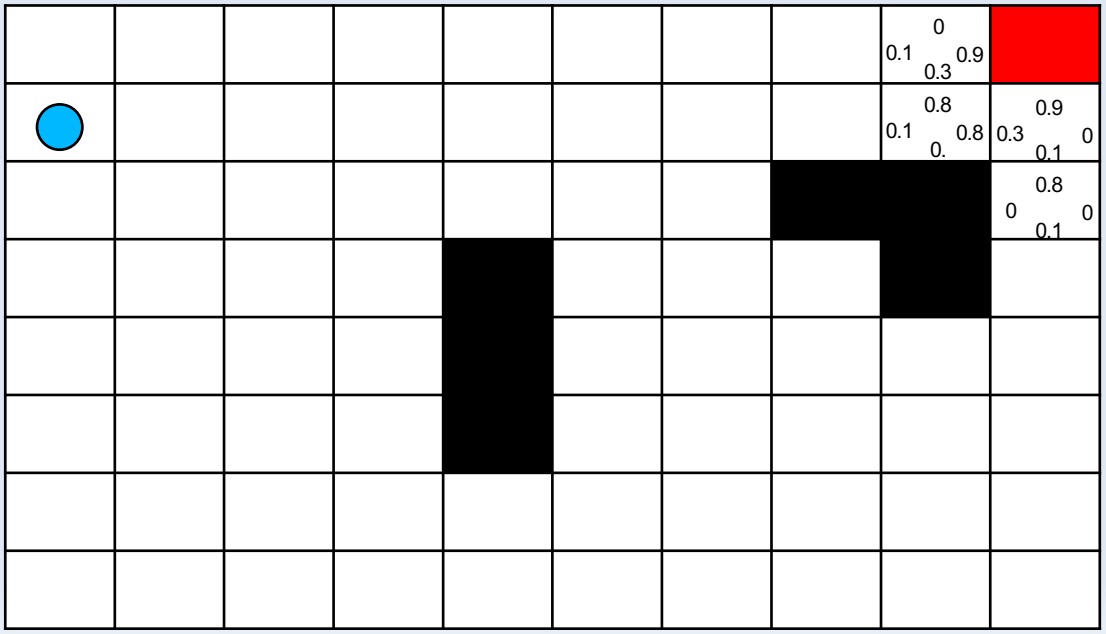
How to learn to perform the right actions

other solution: learning Q-values (qualities)

$Q : (S,A) \longrightarrow R$

Q-table:

state / action	a1 : North	a2 : South	a3 : East	a4 : West
s1	0.92	0.10	0.35	0.05
s2	0.25	0.52	0.43	0.37
s3	0.78	0.9	1.0	0.81
s4	0.0	1.0	0.9	0.9
...	...	...	...	...



The Q-learning model

How can the agent learn a policy?

How to learn to perform the right actions

other solution: learning Q-values (qualities)

$Q : (S,A) \longrightarrow R$

Q-table:

state / action	a1 : North	a2 : South	a3 : East	a4 : West
s1	0.92	0.10	0.35	0.05
s2	0.25	0.52	0.43	0.37
s3	0.78	0.9	1.0	0.81
s4	0.0	1.0	0.9	0.9
...	...	...	...	...

$$P(a) = \frac{\exp(\boldsymbol{\beta} \cdot Q(s,a))}{\Sigma_b \exp(\boldsymbol{\beta} \cdot Q(s,b))}$$

The $\boldsymbol{\beta}$ parameter regulates the exploration – exploitation trade-off.

Different Temporal-Difference (TD) methods

- ACTOR-CRITIC

$$V(s_t) \leftarrow V(s_t) + \alpha[\underbrace{r_{t+1} + \gamma V(s_{t+1}) - V(s_t)}]$$

State-dependent Reward Prediction Error

(independent from the action)

Different Temporal-Difference (TD) methods

- ACTOR-CRITIC

$$V(s_t) \leftarrow V(s_t) + \alpha[r_{t+1} + \gamma V(s_{t+1}) - V(s_t)]$$

- SARSA

$$Q(s_t, a_t) \leftarrow Q(s_t, a_t) + \alpha[r_{t+1} + \gamma Q(s_{t+1}, a_{t+1}) - Q(s_t, a_t)]$$

Reward Prediction Error dependent on the action chosen to be performed next

Different Temporal-Difference (TD) methods

- ACTOR-CRITIC

$$V(s_t) \leftarrow V(s_t) + \alpha[r_{t+1} + \gamma V(s_{t+1}) - V(s_t)]$$

- SARSA

$$Q(s_t, a_t) \leftarrow Q(s_t, a_t) + \alpha[r_{t+1} + \gamma Q(s_{t+1}, a_{t+1}) - Q(s_t, a_t)]$$

- Q-LEARNING

$$Q(s_t, a_t) \leftarrow Q(s_t, a_t) + \alpha[r_{t+1} + \gamma \max_{a \in A} Q(s_{t+1}, a) - Q(s_t, a_t)]$$

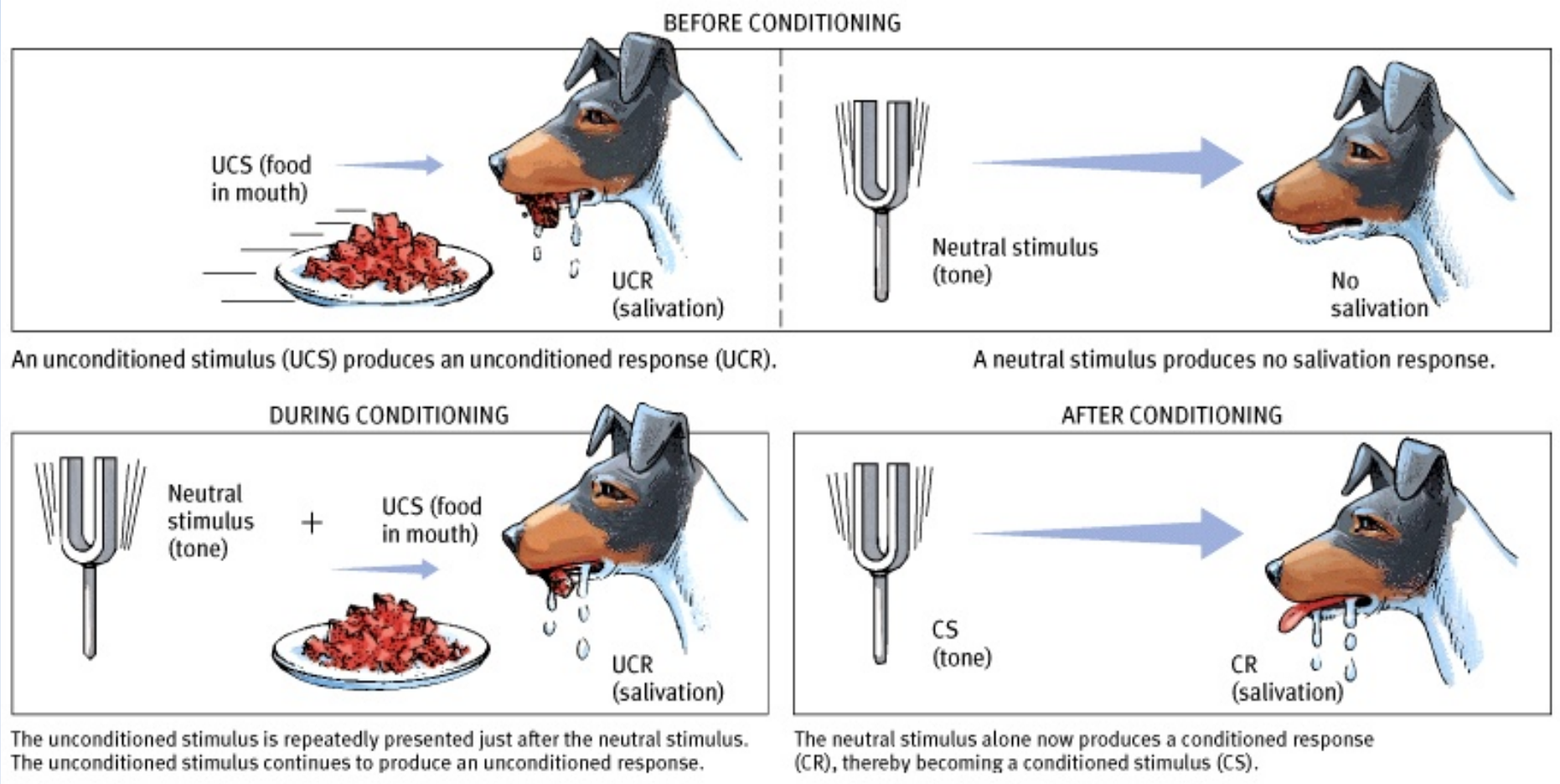
Reward Prediction Error dependent on the best action

Links with biology

Activity of dopaminergic neurons

CLASSICAL CONDITIONING

TD-learning explains classical conditioning (predictive learning)



Taken from Bernard Balleine's lecture at Okinawa Computational Neuroscience Course (2005).

REINFORCEMENT LEARNING

- Analogy with dopaminergic neurons' activity

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

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

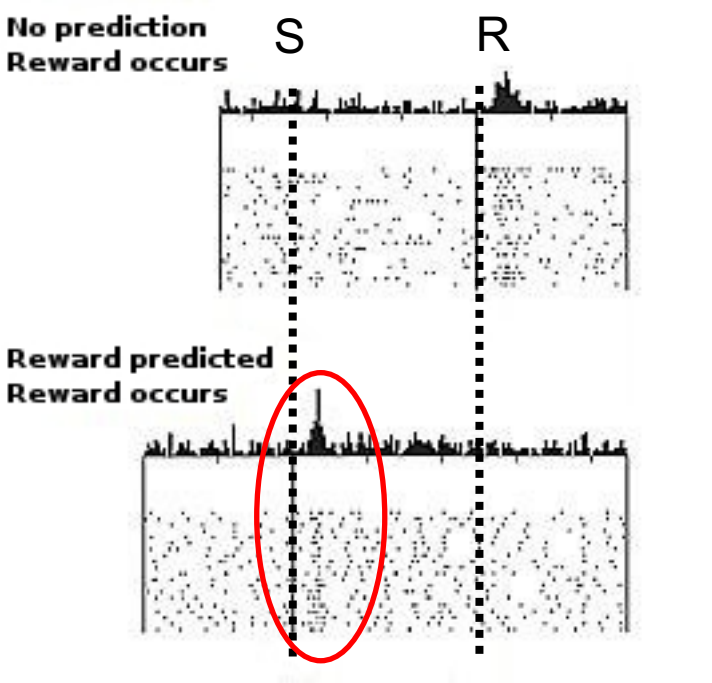
+1

reinforcement

reward

REINFORCEMENT LEARNING

- Analogy with dopaminergic neurons' activity



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

+1

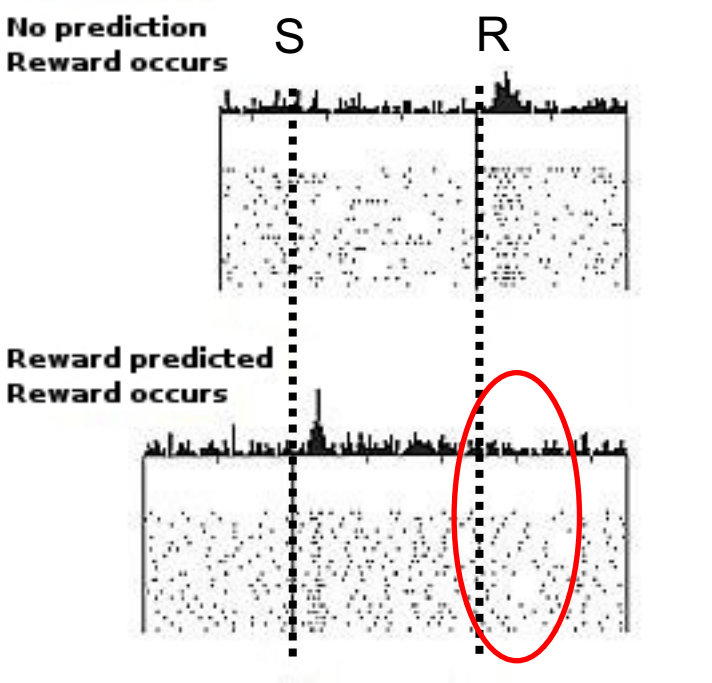
$$\delta_{t+1} = r_{t+1} + \gamma \cdot V(s_{t+1}) - V(s_t)$$

reinforcement

reward

REINFORCEMENT LEARNING

- Analogy with dopaminergic neurons' activity



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

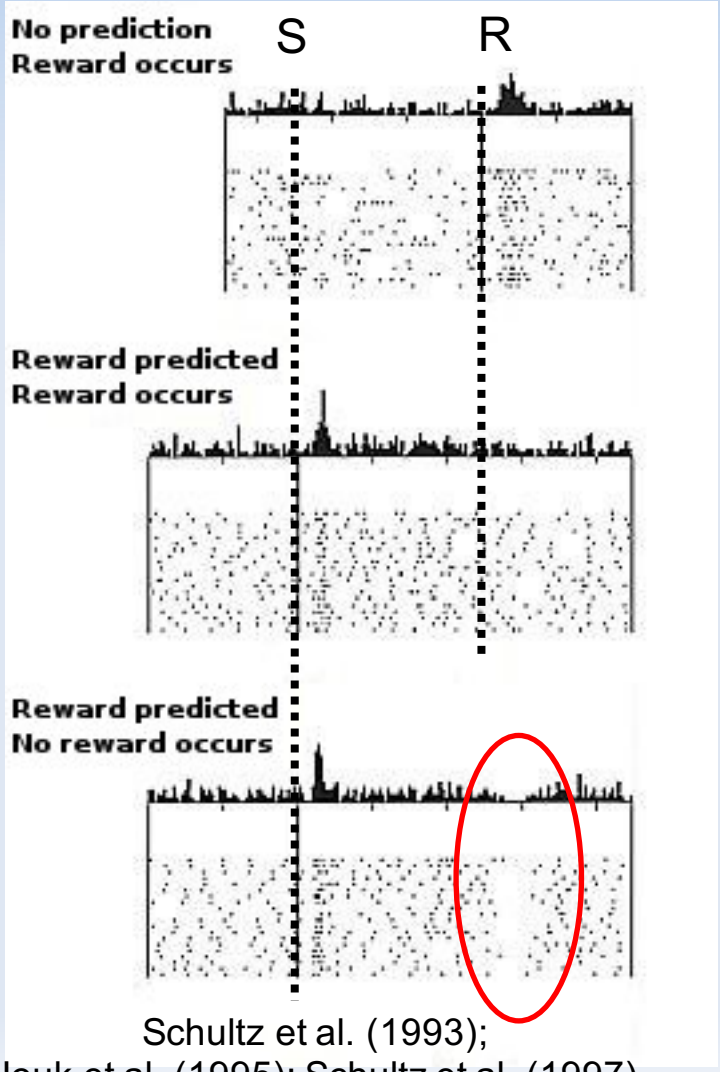
$$\delta_{t+1} = r_{t+1} + \gamma \cdot \underbrace{V(s_{t+1})}_{0} - V(s_t)$$

reinforcement

reward

REINFORCEMENT LEARNING

- Analogy with dopaminergic neurons' activity



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

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

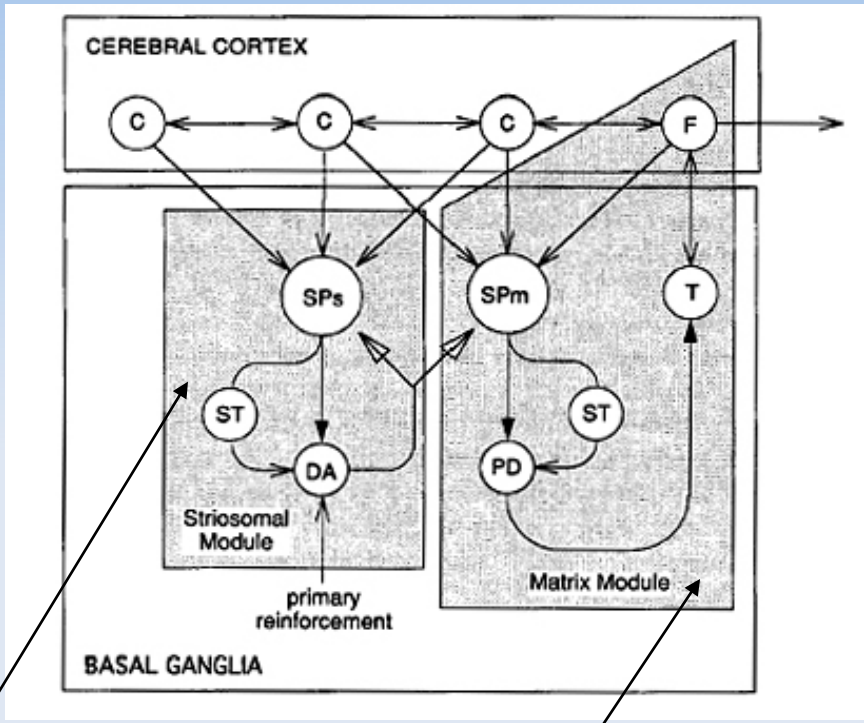
-1

reinforcement

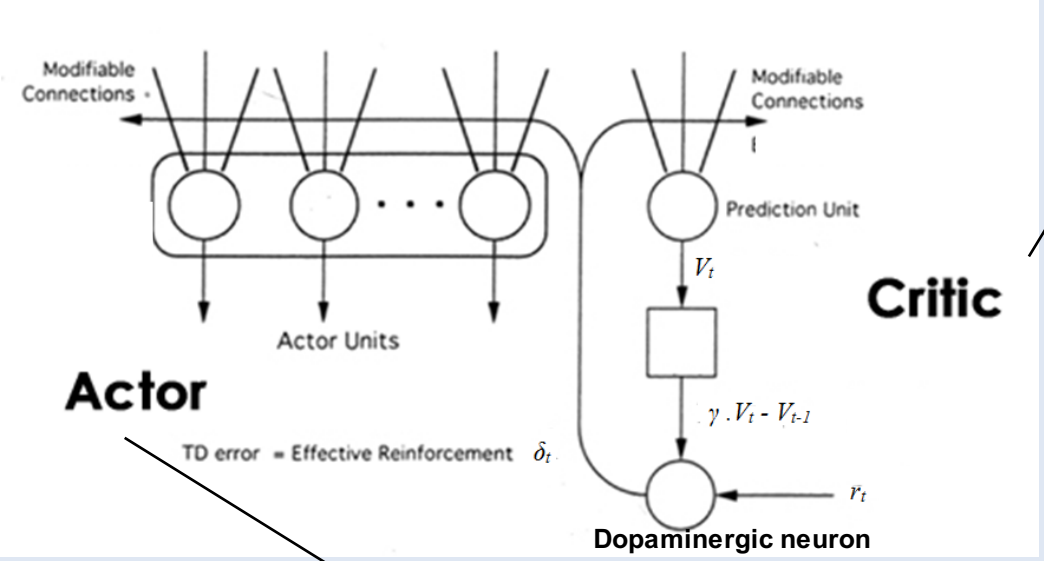
reward

The Actor-Critic model and the Basal Ganglia

Barto (1995); Montague et al. (1996); Schultz et al. (1997); Berns and Sejnowski (1996); Suri and Schultz (1999); Doya (2000); Suri et al. (2001); Baldassarre (2002).
see Joel et al. (2002) for a review.



Houk et al. (1995)



Wide application of RL models to model-based analyses of behavioral and physiological data during decision-making tasks

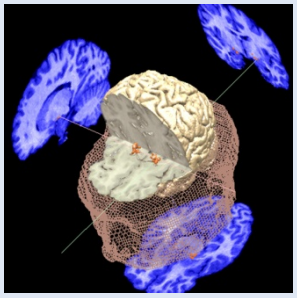
Model-based analysis of brain data

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

fMRI scanner

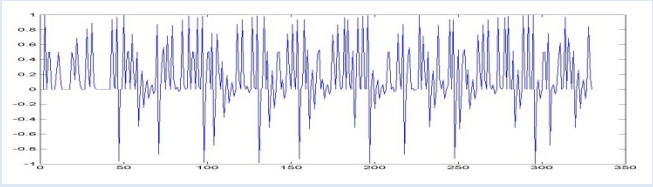
RL model

Brain responses



?

Prediction error



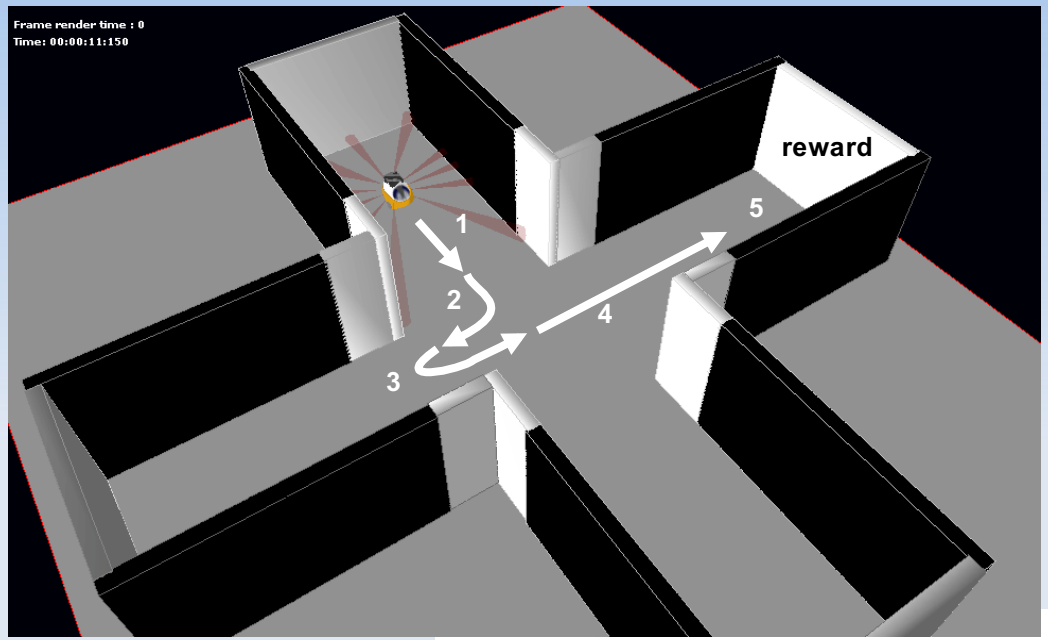
cf. travail de Mathias Pessiglione (ICM)

ou Giorgio Coricelli (ENS)

If we can find reward prediction error signals, do we also find reward predicting signals?

→ REWARD PREDICTION IN THE STRIATUM

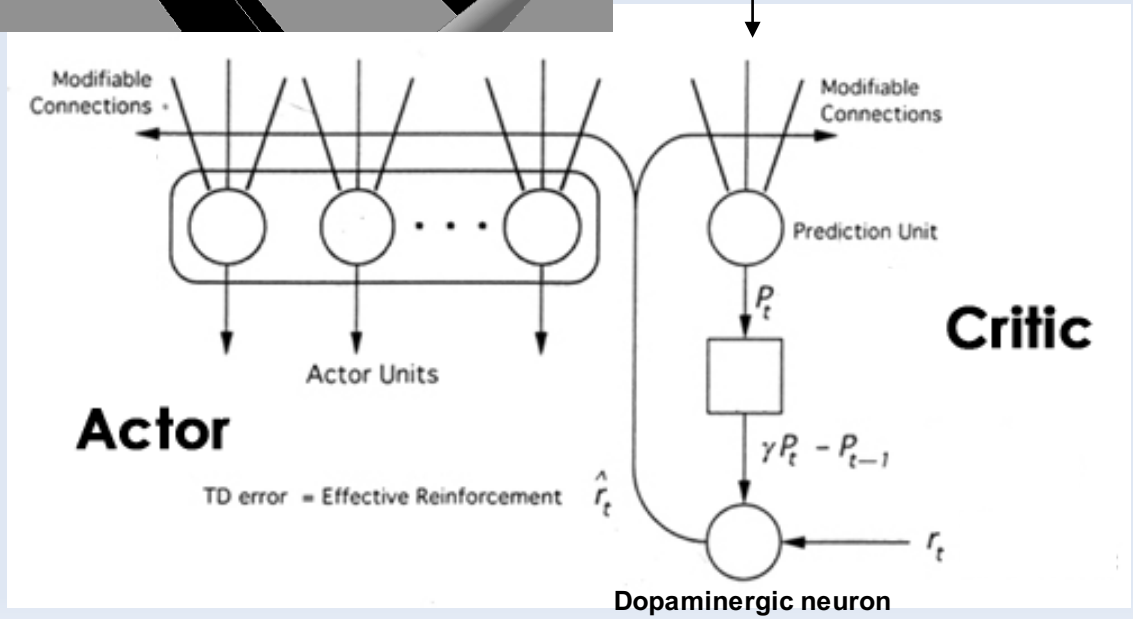
The Actor-Critic model



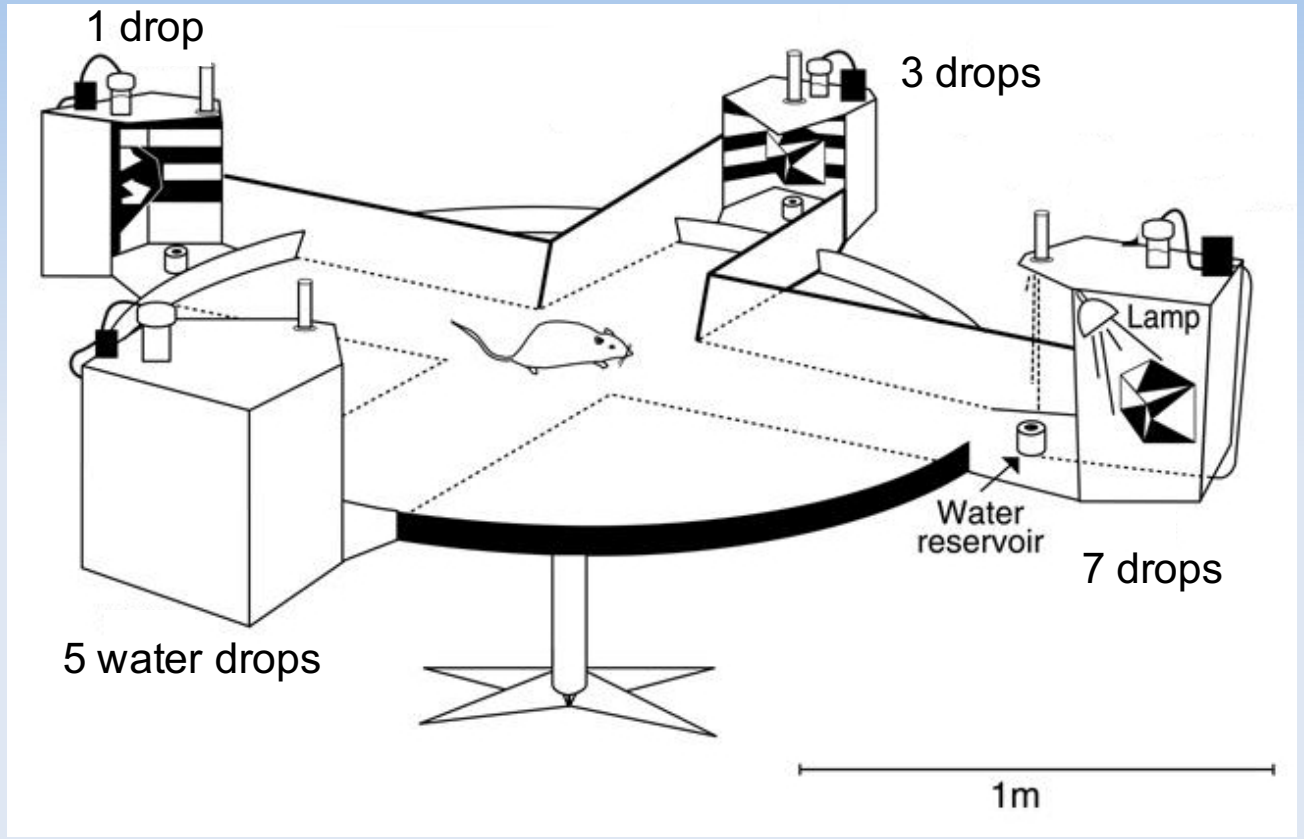
Which state space as an input?

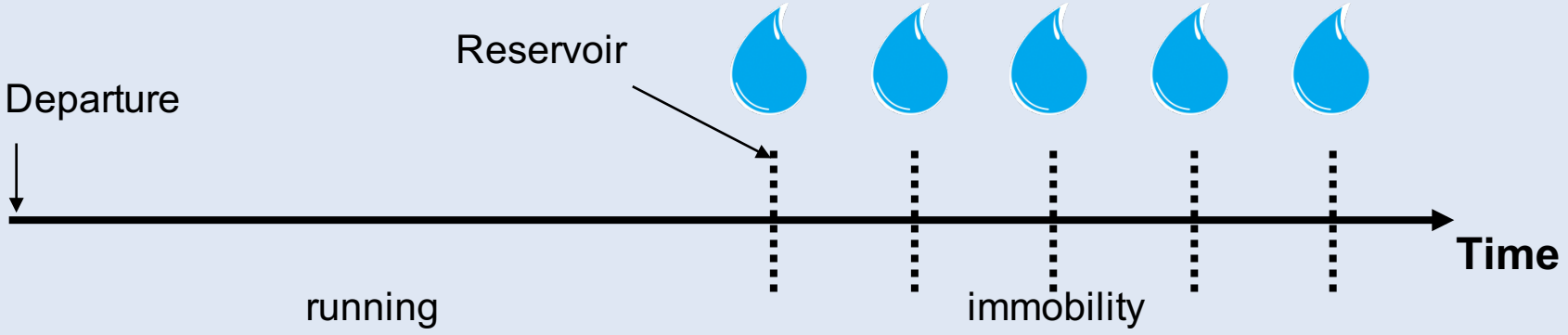
Temporal-order input
[0 0 1 0 0 0 0]

or spatial or visual information

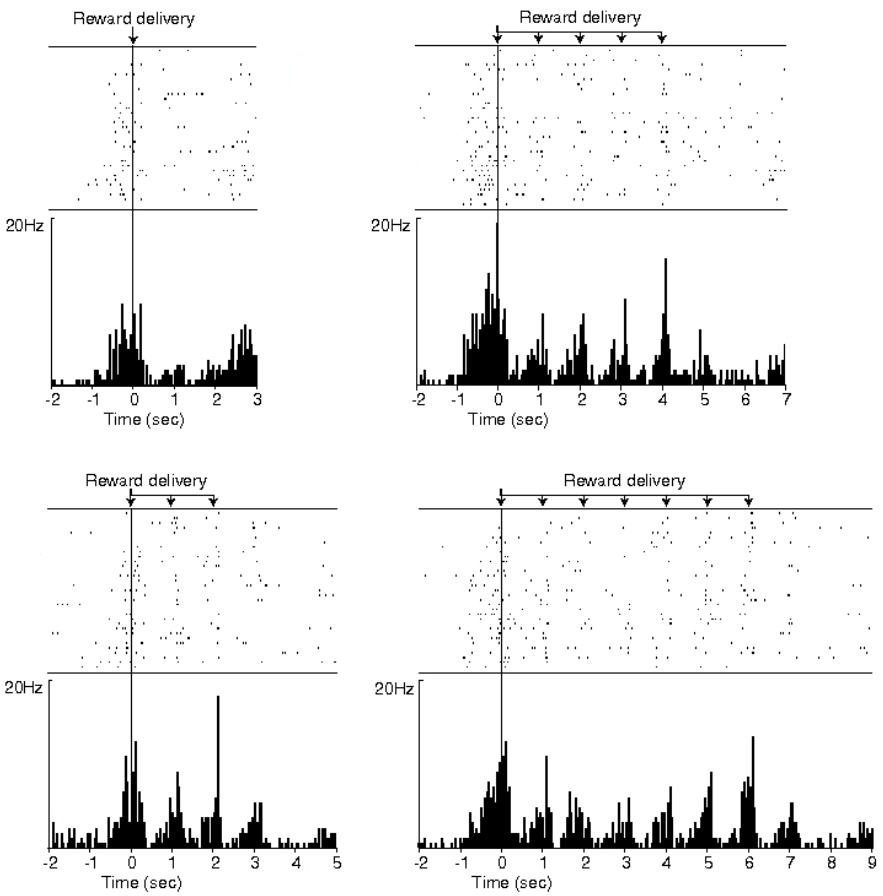


Electrophysiology
Reward prediction in the striatum

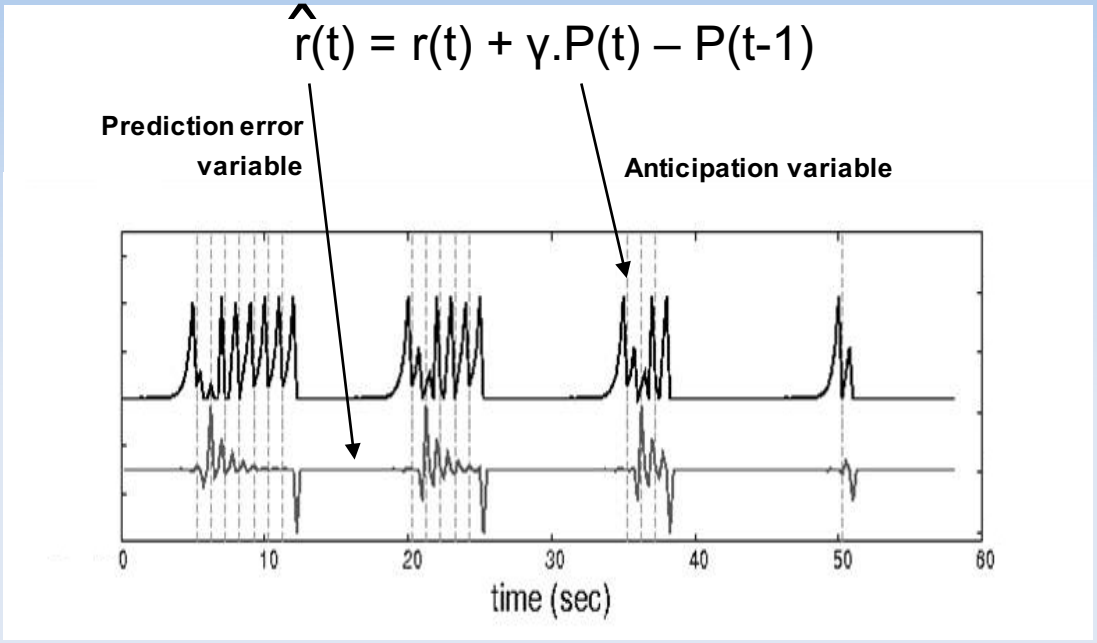




RESULTS: Coherent with the TD-learning model



Activity of a neuron from striatum



Simulated TD-learning model

Corrélés

Khamassi, Mulder, Tabuchi, Douchamps & Wiener (2008). European Journal of Neuroscience.

Modelling with TD-learning Results

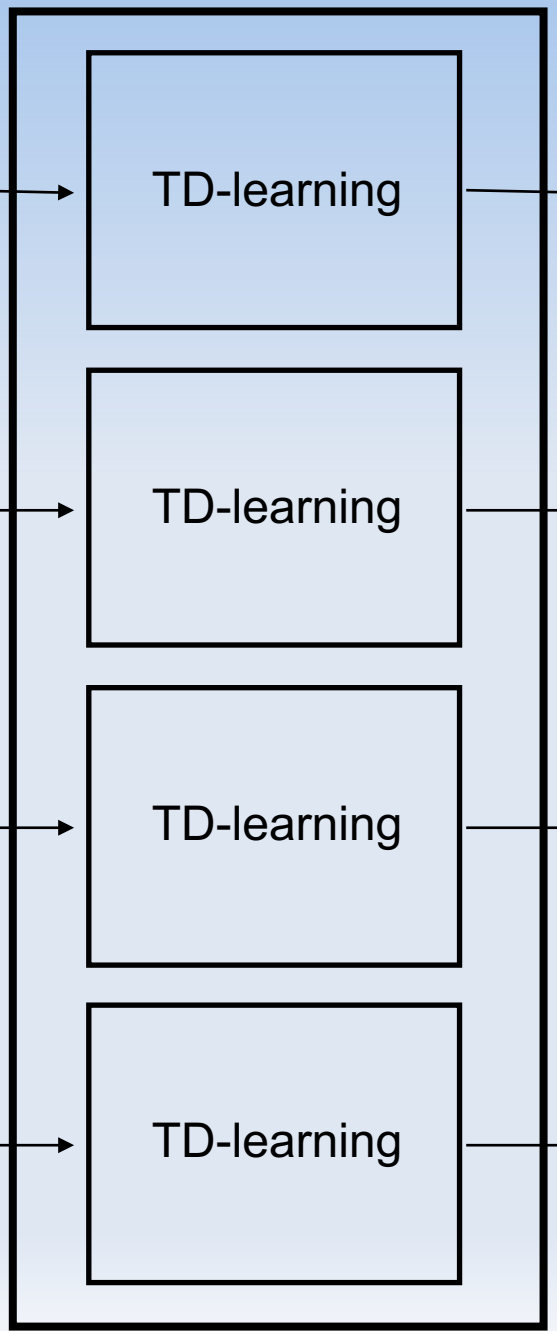
Temporal order information
(Montague et al., 1996).
[0 0 1 0 0] [0 0 0 0 0] ...

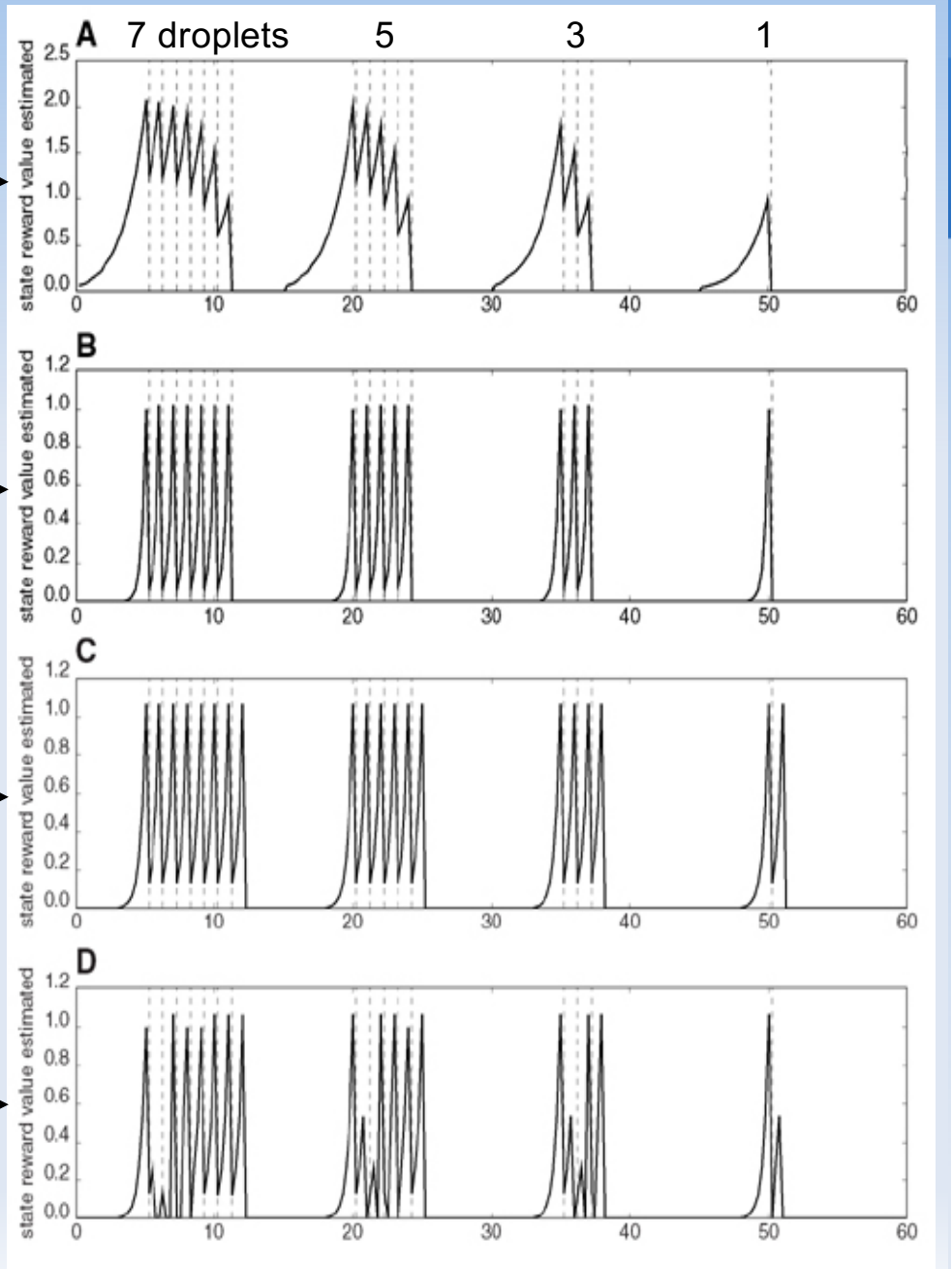
Incomplete temporal representation
[0 0 1] [0 0 0] ...

Ambiguous visual input
[0 0 1] [0 0 0] ...

No spatial information
[0 0 1] [0 0 1] ...

Place #1 Place #2





This works well, but…

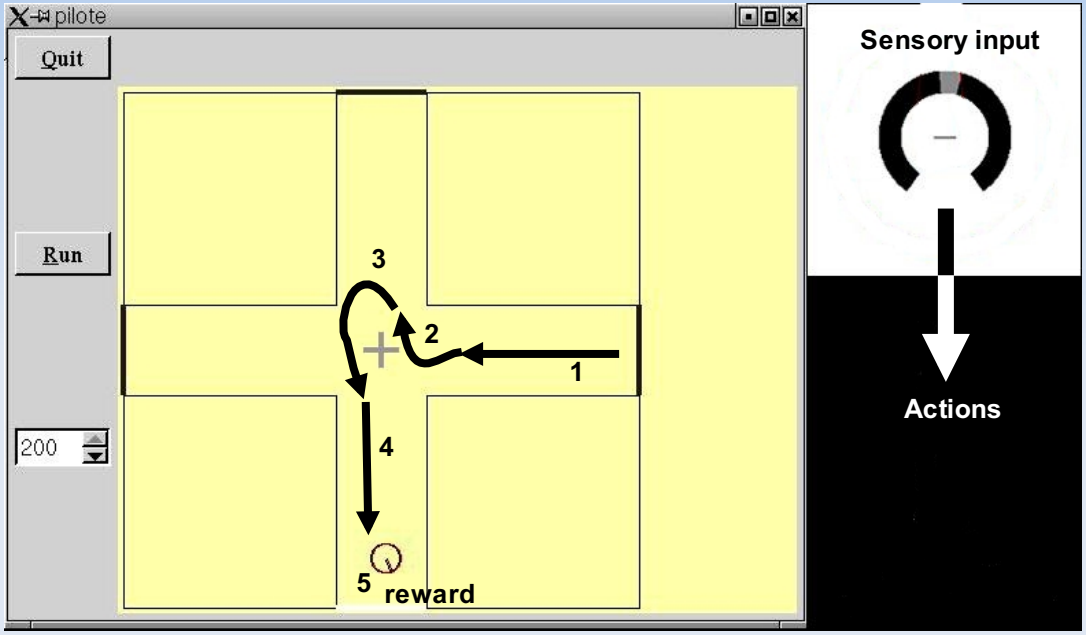
- *Most experiments are single-step*
- *All these cases are discrete*
- *Very small number of states, actions*
- *We supposed a perfect state identification*

OUTLINE

1. Intro
2. Reinforcement Learning model
   - Algorithm
   - Dopamine activity
3. Continuous RL
   - Robot navigation
   - Neuro-inspired models
4. PFC & off-line learning
   - Indirect reinforcement learning
   - Replay during sleep
5. Meta-Learning
   - Principle
   - Neuronal recordings
   - Humanoid Robot interaction

CONTINUOUS REINFORCEMENT LEARNING

Robotics application

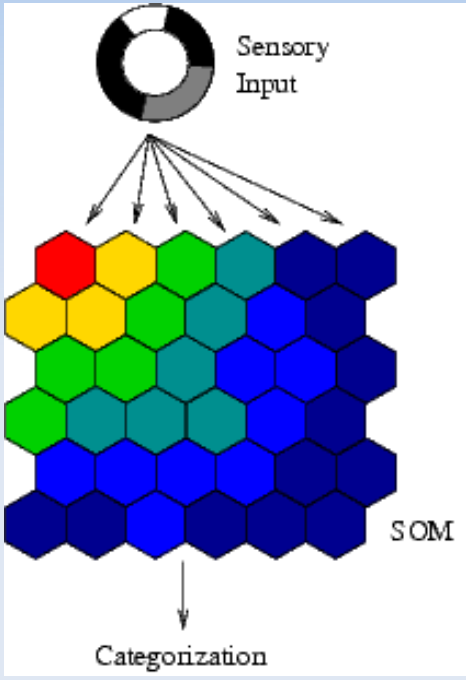


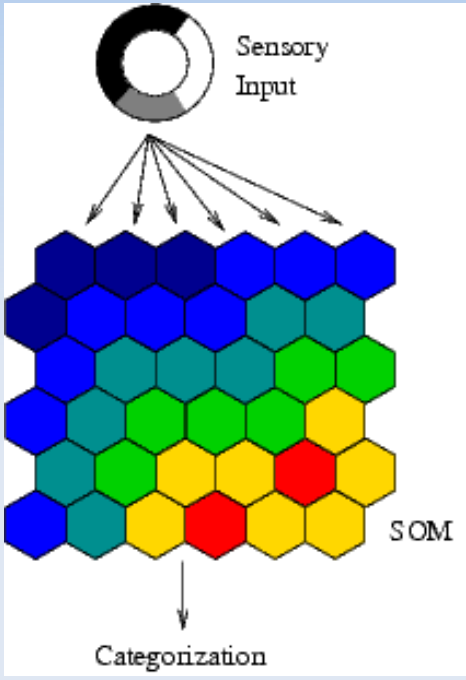
TD-Learning model applied to spatial navigation behavior learning in the plus-maze task

Khamassi et al. (2005). *Adaptive Behavior.*
Khamassi et al. (2006). *Lecture Notes in Computer Science*

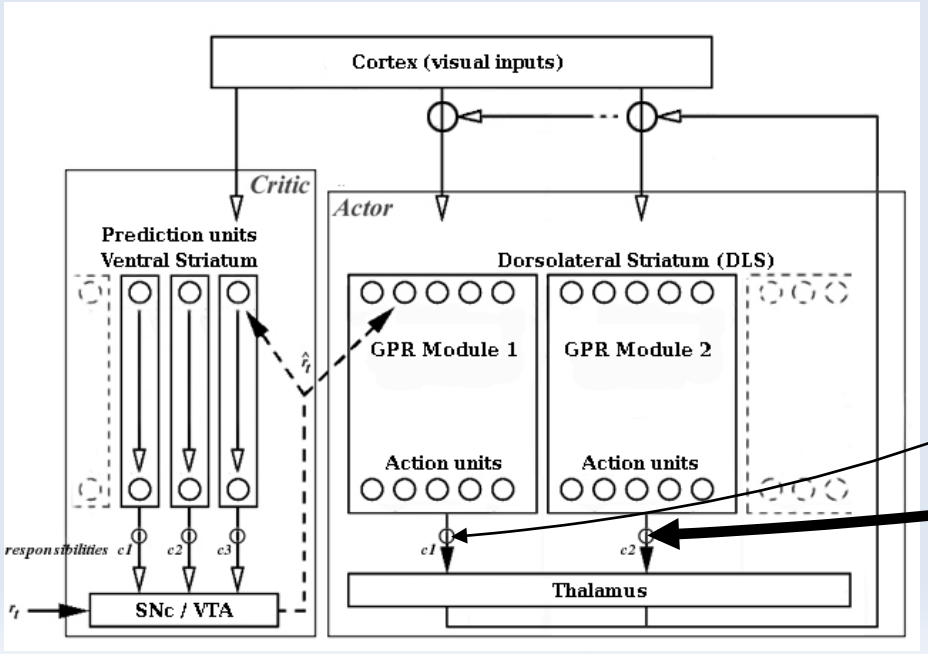
Extension of the Actor-Critic model

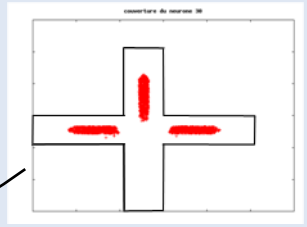
Coordination by a self-organizing map

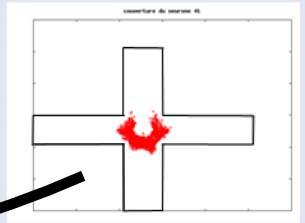




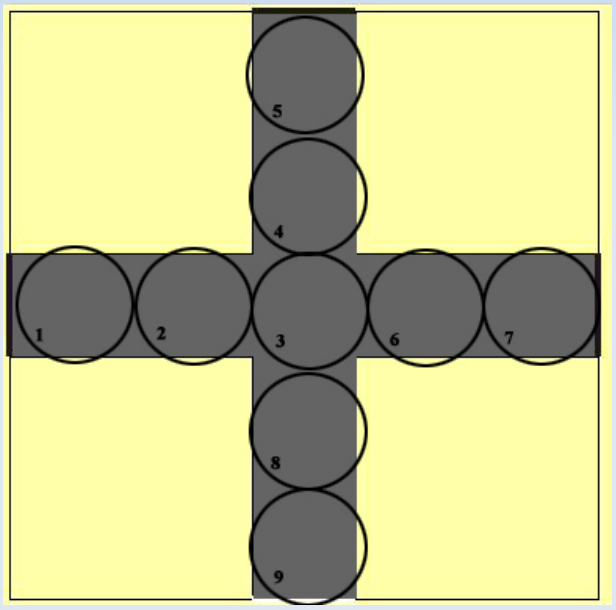
Actor-Critic multi-modules neural network



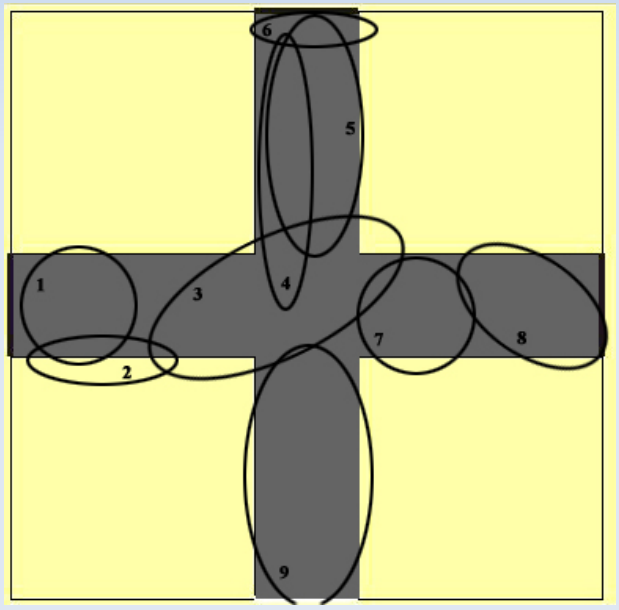




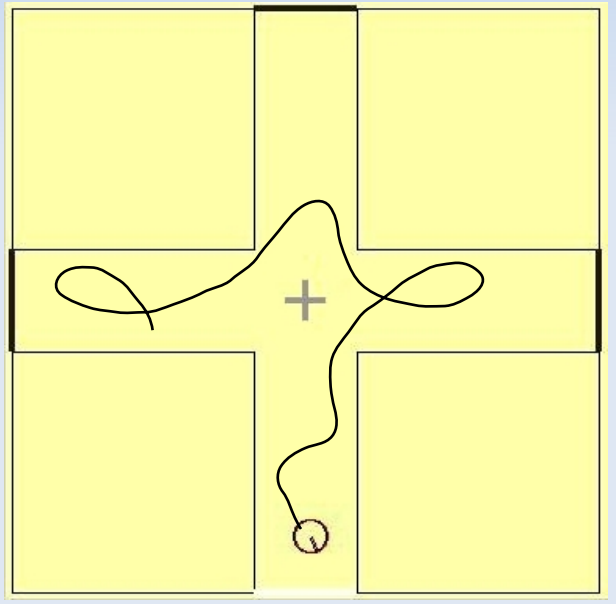
Extension of the Actor-Critic model



Hand-tuned

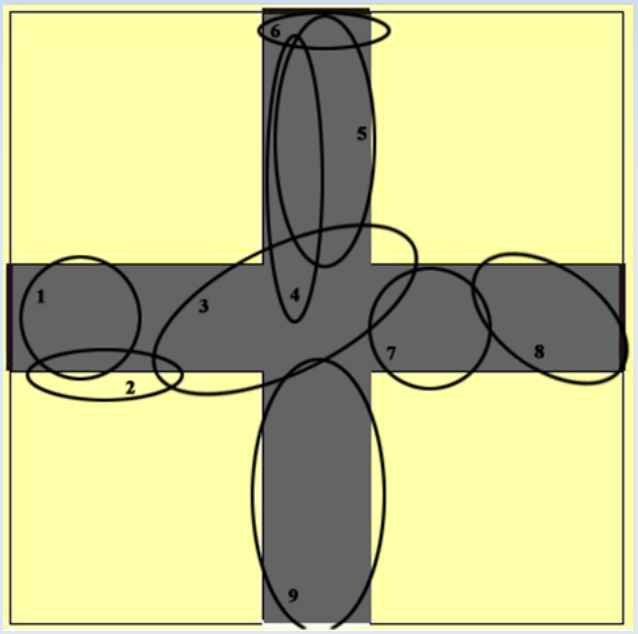


Autonomous



Random

Extension of the Actor-Critic model



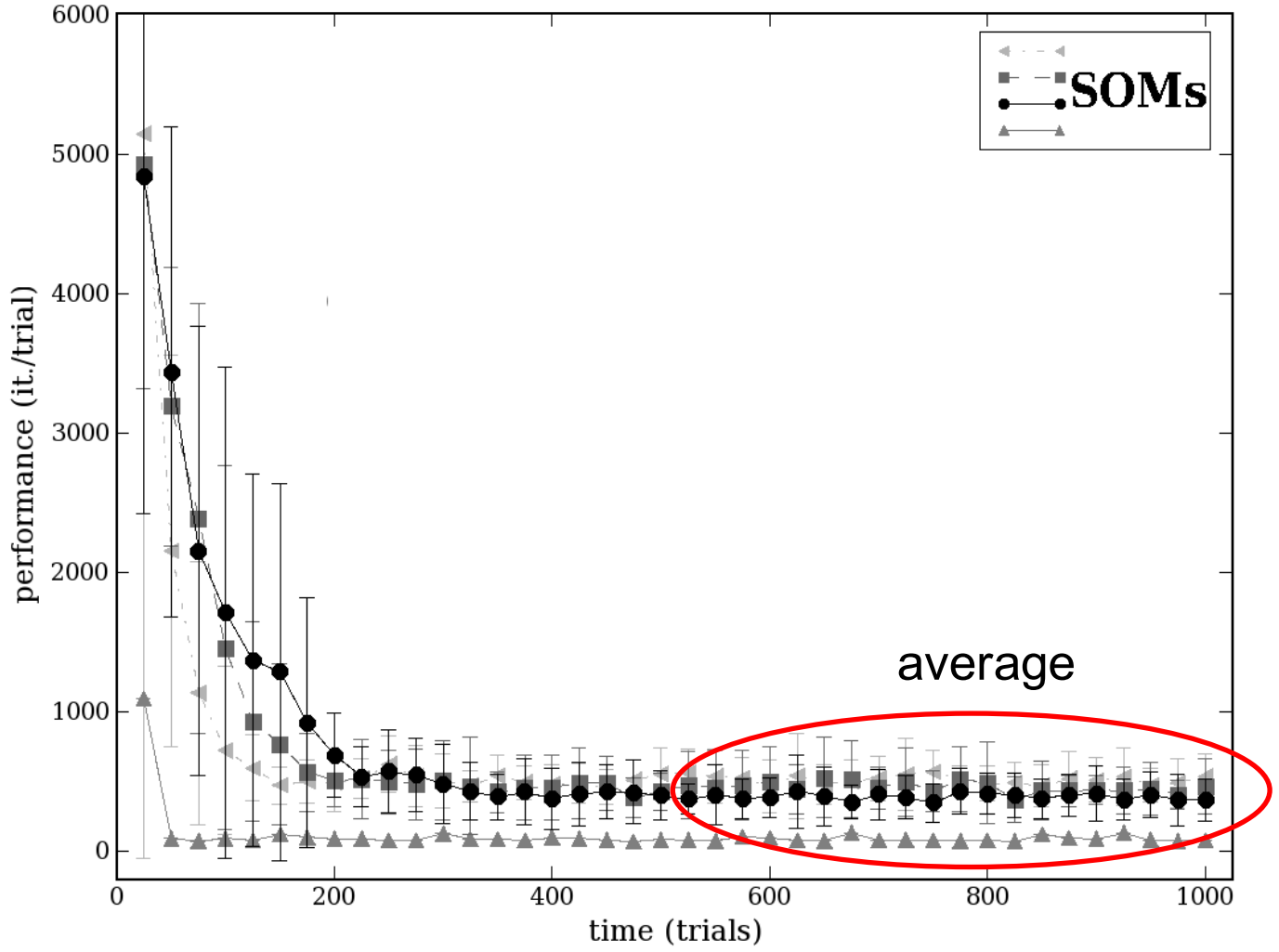
Autonomous

Two methods :

1. Self-Organizing Maps (SOMs)

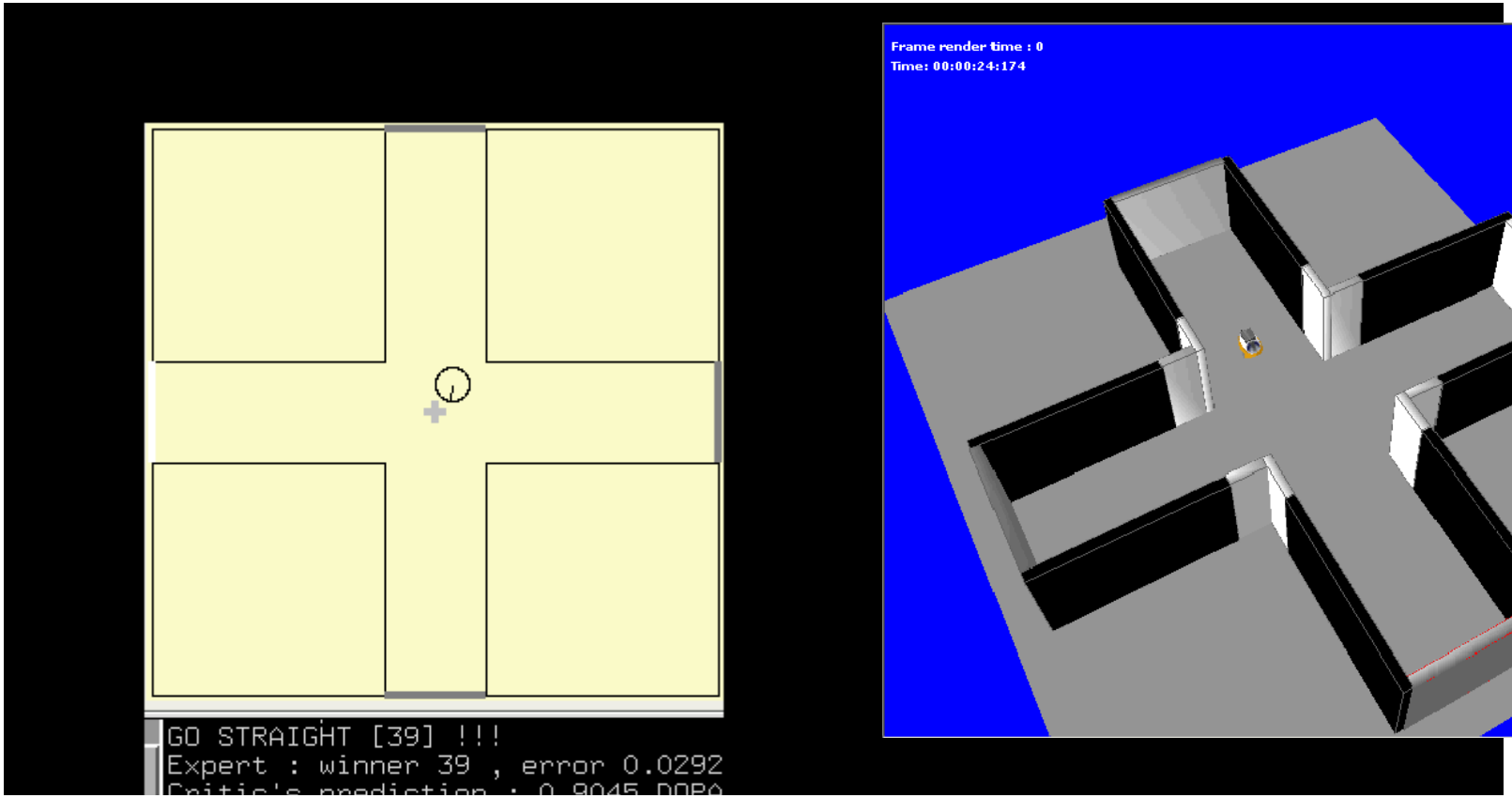
2. specialization based on performance (tests modules' capacity for state prediction) Baldassarre (2002); Doya et al. (2002). Within a particular subpart of the maze, only the module with the most accurate reward prediction is trained. Each module thus becomes an expert responsible for learning in a given task subset.

Extension of the Actor-Critic model



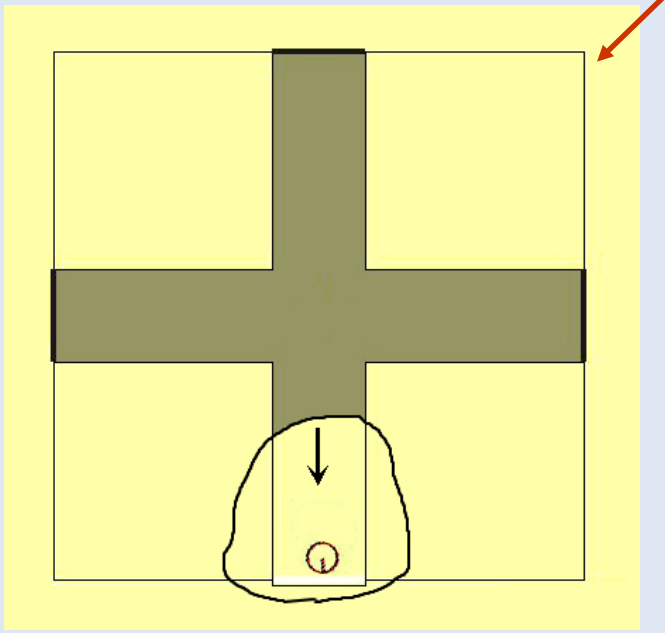
Extension of the Actor-Critic model

	Nb of iterations required (Average performance during the second half of the experiment)
1. hand-tuned	94
2. specialization based on performance	3,500
3. autonomous categorization (SOM)	404
4. random robot	30,000



Extension of the Actor-Critic model

	Nb of iterations required (Average performance during the second half of the experiment)
1. hand-tuned	94
2. specialization based on performance	3,500
3. autonomous categorization (SOM)	404
4. random robot	30,000



OUTLINE

1. Intro
2. Reinforcement Learning model
   - Algorithm
   - Dopamine activity
3. Continuous RL
   - Robot navigation
   - Neuro-inspired models
4. PFC & off-line learning
   - Indirect reinforcement learning
   - Replay during sleep
5. Meta-Learning
   - Principle
   - Neuronal recordings
   - Humanoid Robot interaction

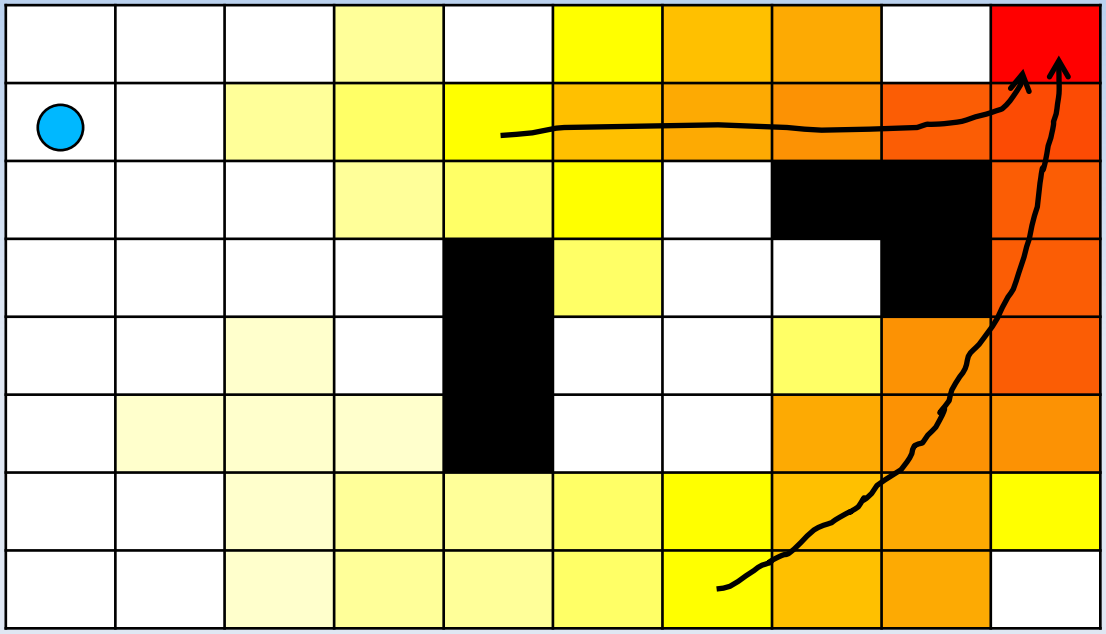
Off-learning (Indirect RL) & prefrontal cortex activity during sleep

REINFORCEMENT LEARNING

After

N simulations

Very long!



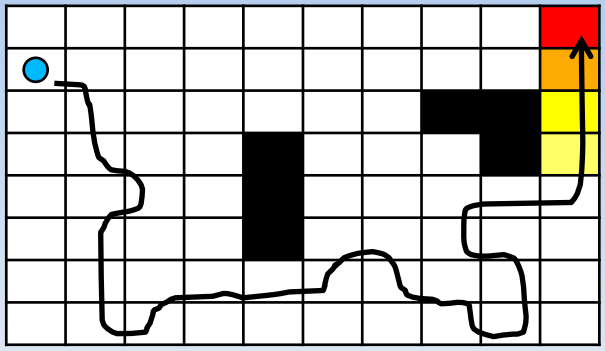
$$\delta_{t+1} = r_{t+1} + \gamma \,.\, V(s_{t+1}) - V(s_t)$$

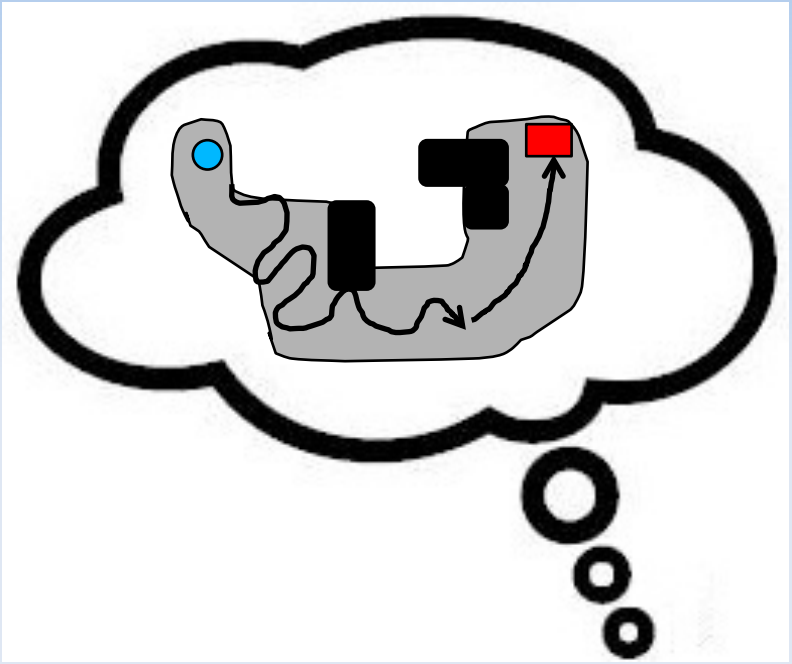
discount factor (=0.9)

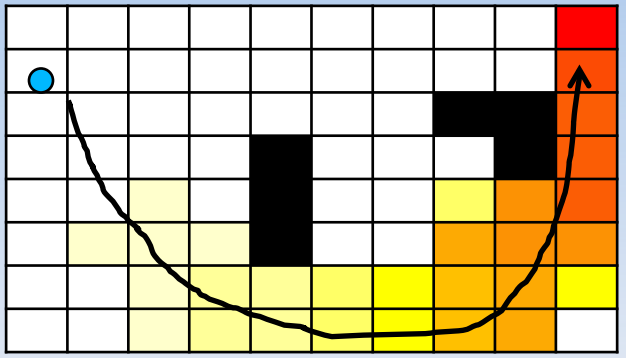
$$V(s_t) = V(s_t) + \alpha \,.\, \delta_{t+1}$$

learning rate (=0.1)

TRAINING DURING SLEEP







Method in Artificial Intelligence:
Off-line Dyna-Q-learning
(Sutton & Barto, 1998)

Model-based Reinforcement Learning

To incrementally learn a model of transition and reward functions, then plan within this model by updates "in the head of the agent" (Sutton, 1990).

$\boldsymbol{S}$: state space

$\boldsymbol{A}$: action space

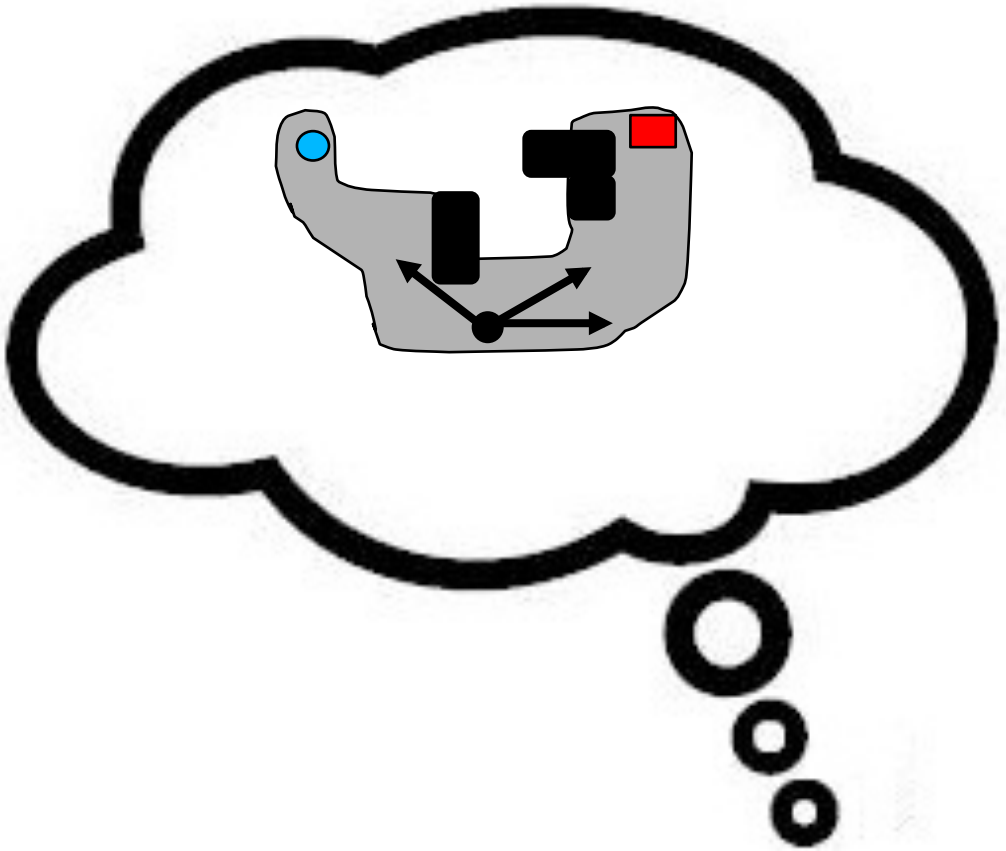
Transition function $\boldsymbol{T} : S \times A \longrightarrow S$

Reward function $\boldsymbol{R} : S \times A \longrightarrow R$

} Internal model

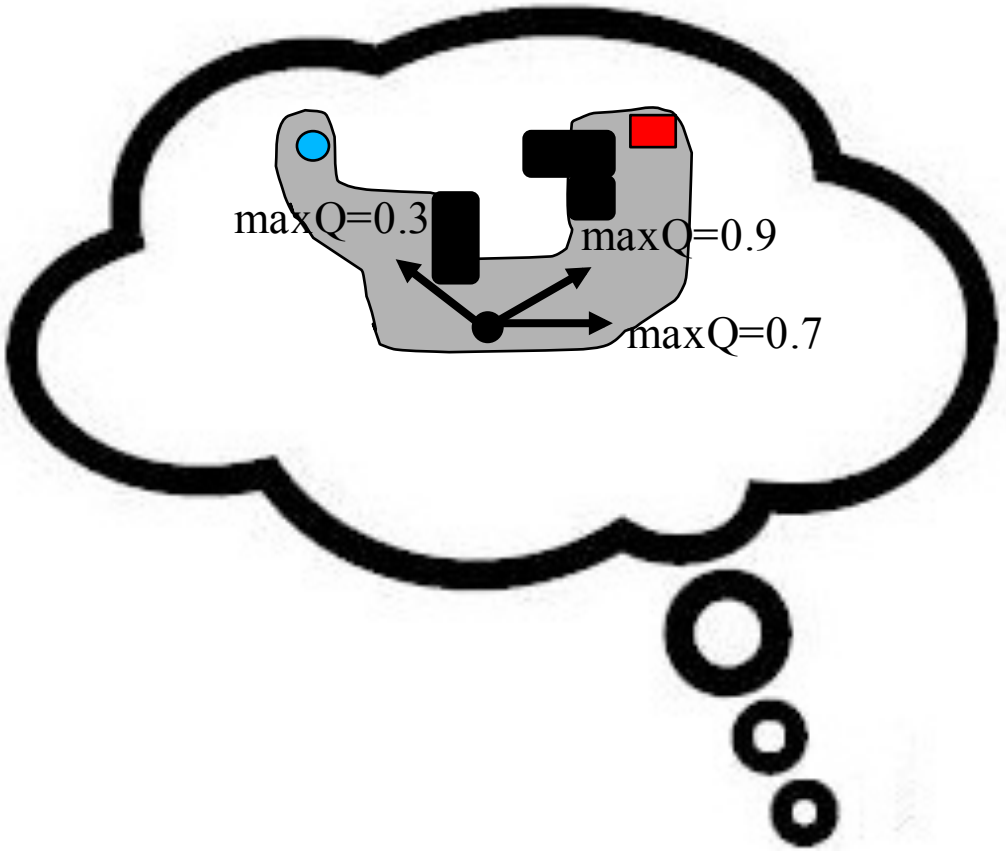
Model-based Reinforcement Learning

s : state of the agent (●)



Model-based Reinforcement Learning

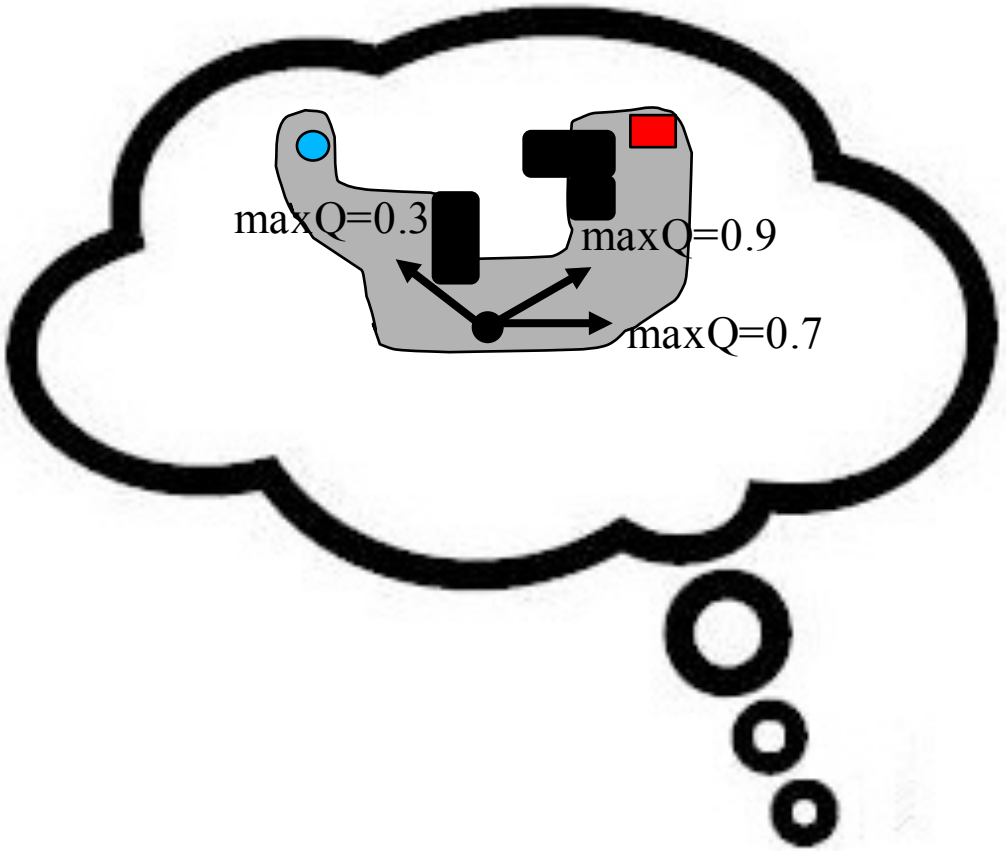
s : state of the agent (●)



Model-based Reinforcement Learning

s : state of the agent (●)

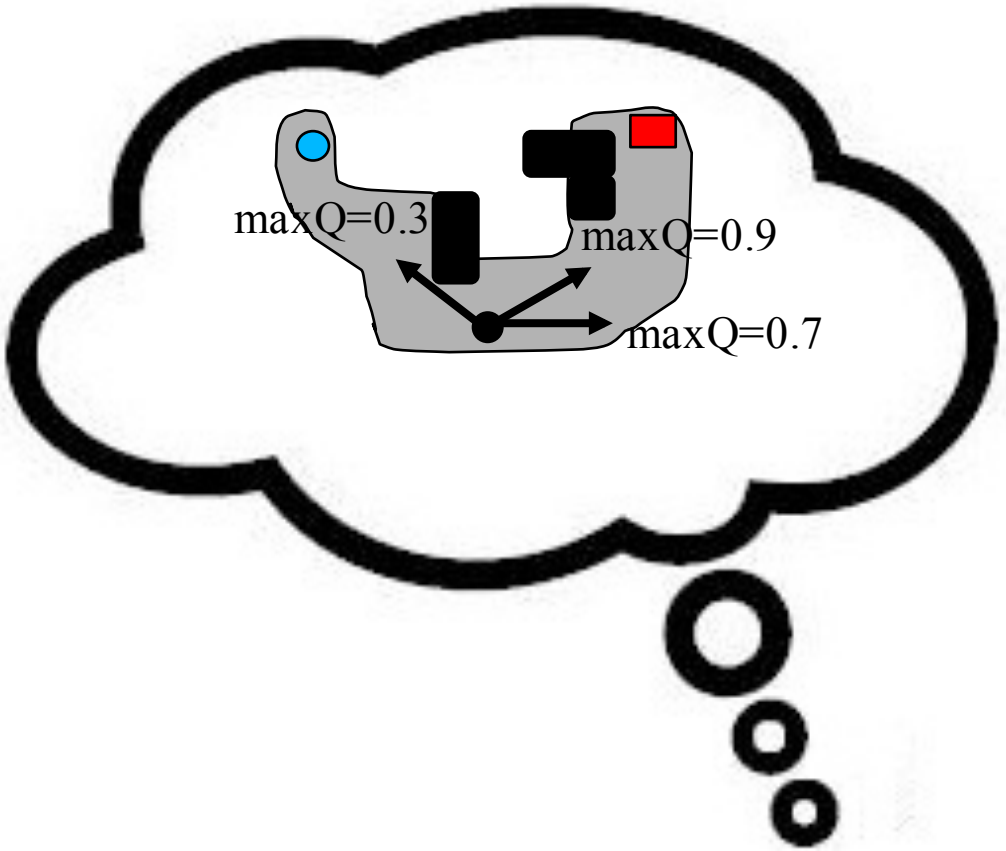
a : action of the agent (**go east**)



Model-based Reinforcement Learning

s : state of the agent (●)

a : action of the agent (**go east**)



stored transition function T:

proba(→) = 0.9

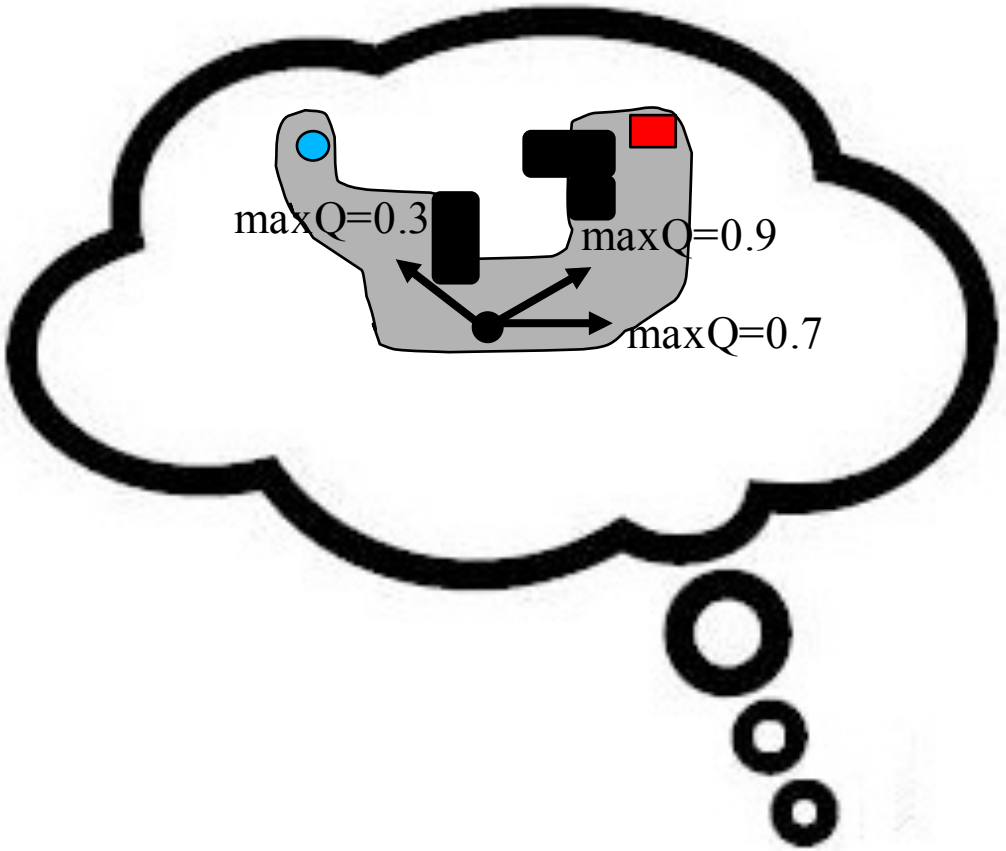
proba(↗) = 0.1

proba(↖) = 0

Model-based Reinforcement Learning

s : state of the agent (●)

a : action of the agent (**go east**)



stored transition function T:

proba(→ *) = 0.9*

proba(↗ *) = 0.1*

proba(↖ *) = 0*

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

0.6 0 0.9*0.7 + 0.1*0.9 + 0*0.3 + …

Model-based Reinforcement Learning

No reward prediction error!

Only:

Estimated Q-values

Transition function

Reward function

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

Links with Neuroscience data

Instrumental conditioning (Daw et al., 2005)

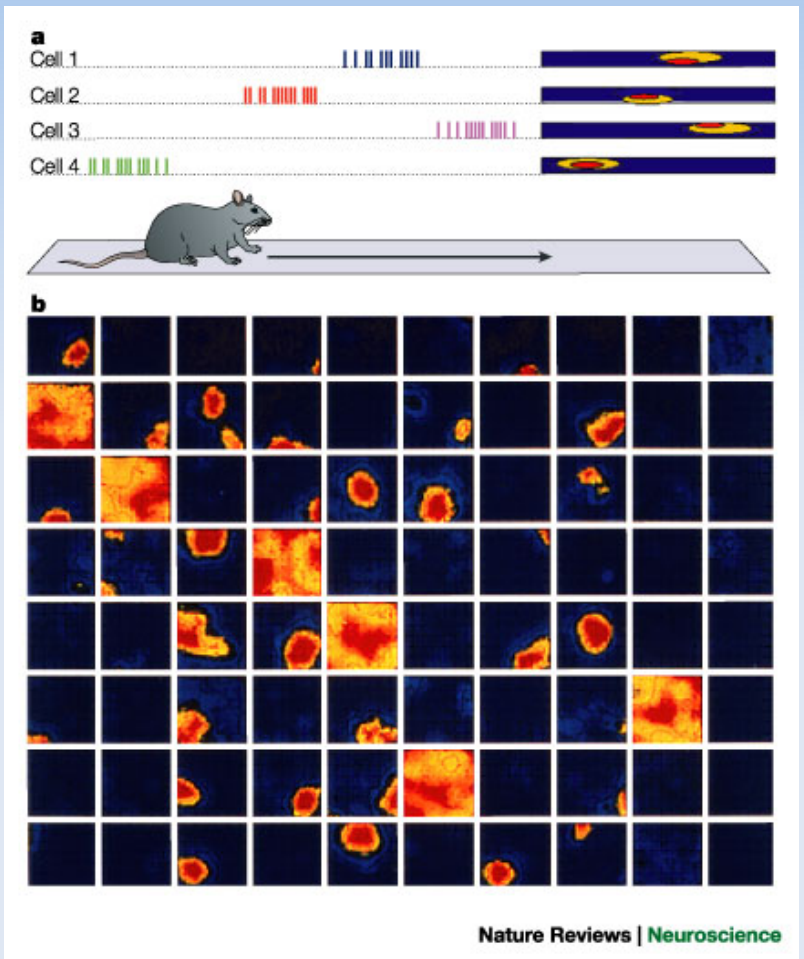
Human behavior (Daw et al., 2011)

Hippocampal off-line replays… (Foster & Wilson, 2006; Euston et al., 2007; Gupta et al., 2010)

…coordinated with PFC or VS (Lansink et al., 2009; Peyrache et al., 2009; Benchenane et al., 2010).

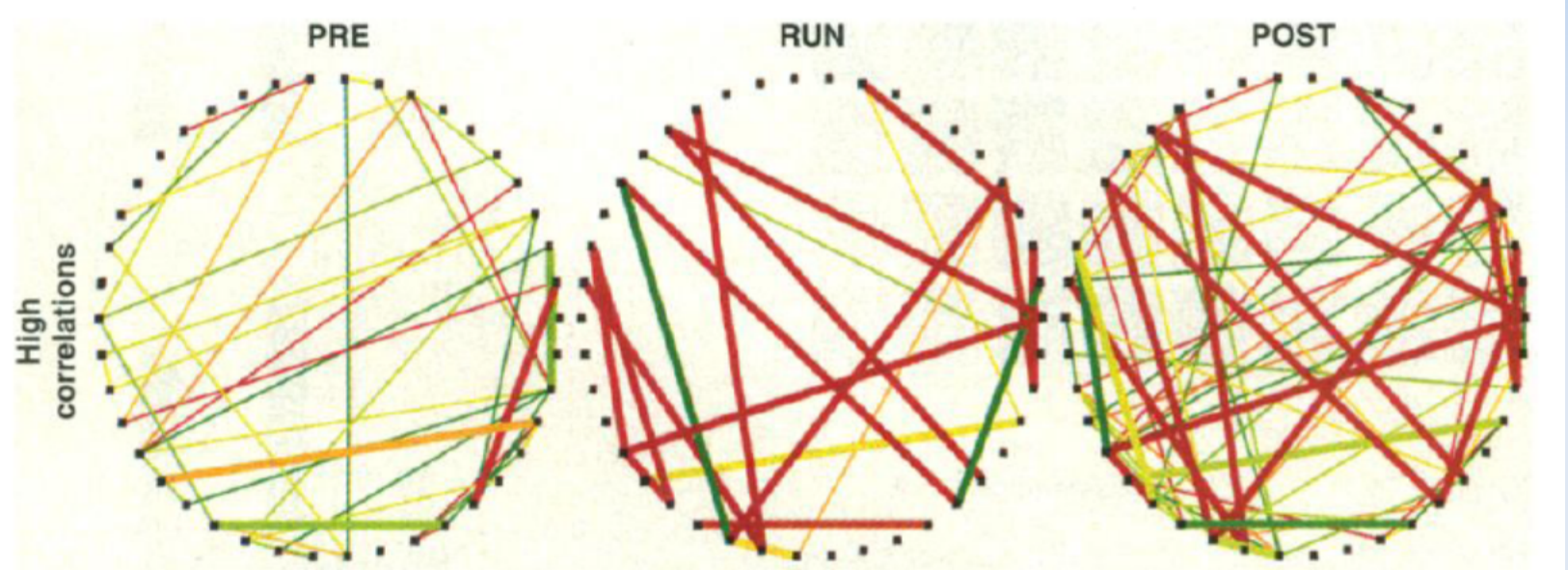
Navigation strategies (Khamassi & Humphries, 2012)

Hippocampal place cells



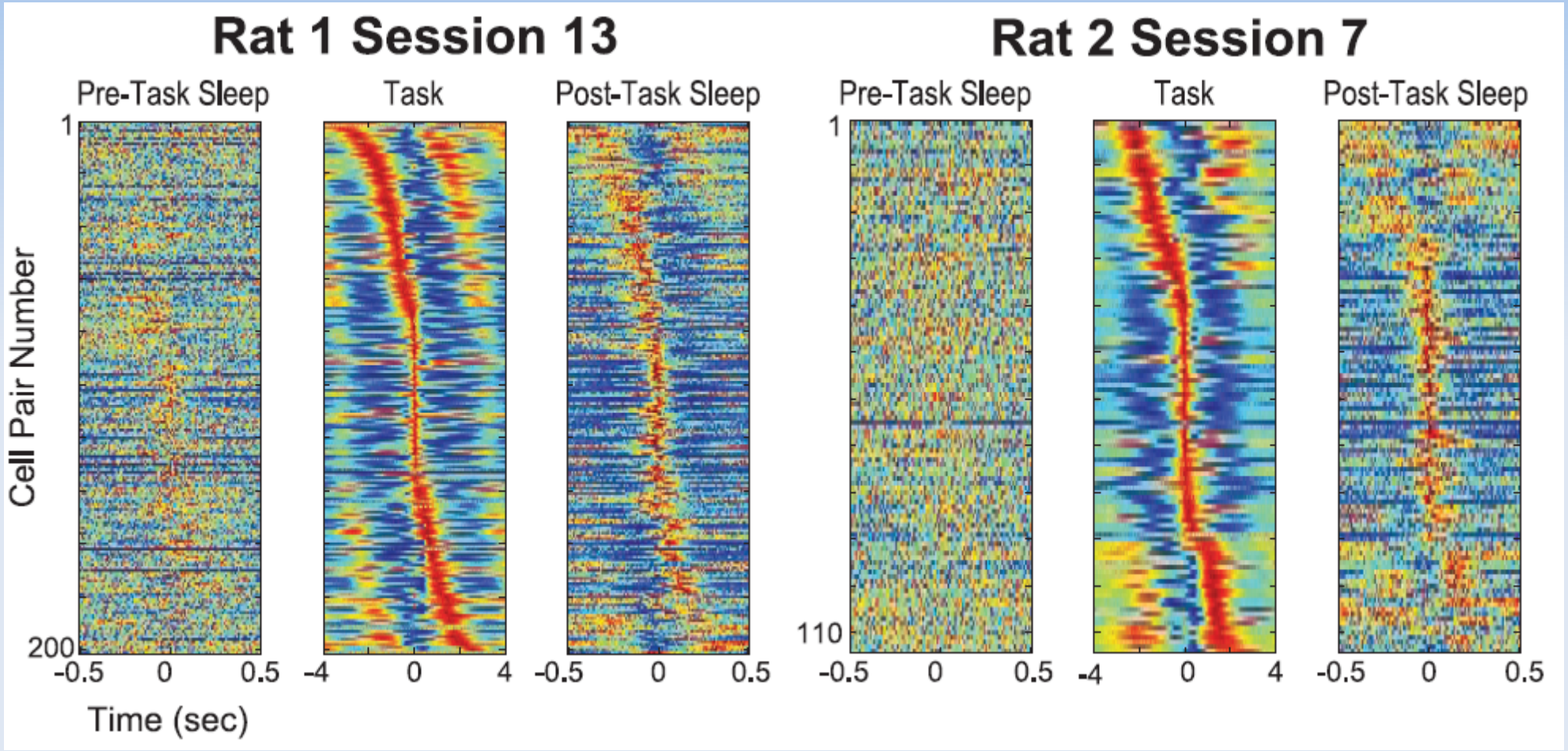
- NMDA receptors, place cells and hippocampal spatial memory. Kazu Nakazawa, Thomas J. McHugh, Matthew A. Wilson & Susumu Tonegawa. Nature Reviews Neuroscience 5, 361-372 (May 2004)

Hippocampal place cells



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

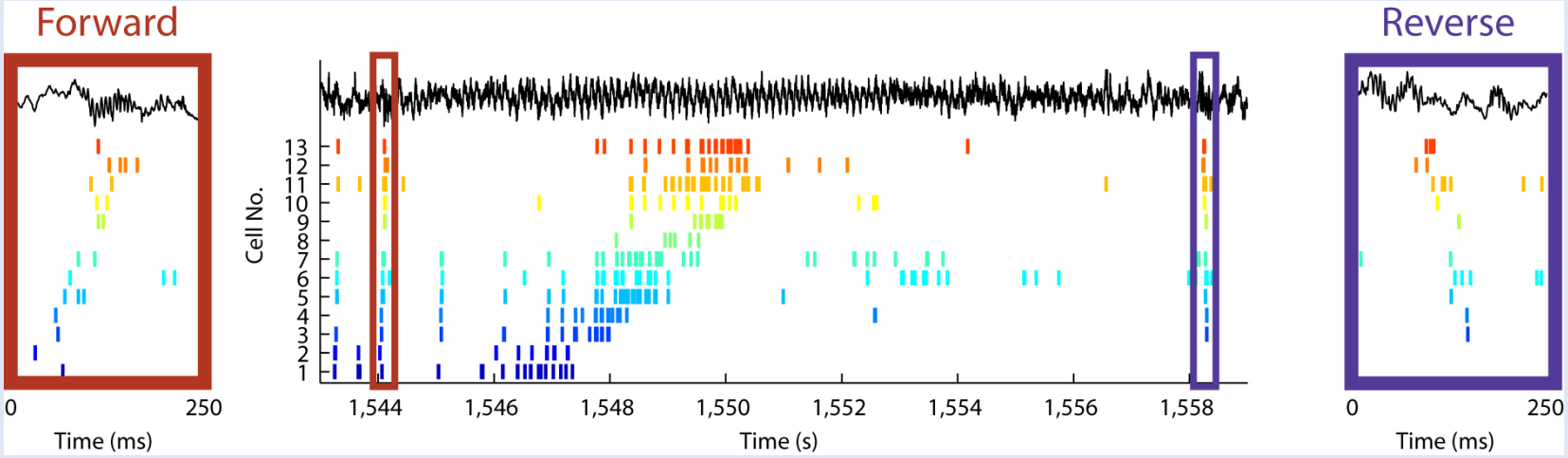
Hippocampal place cells



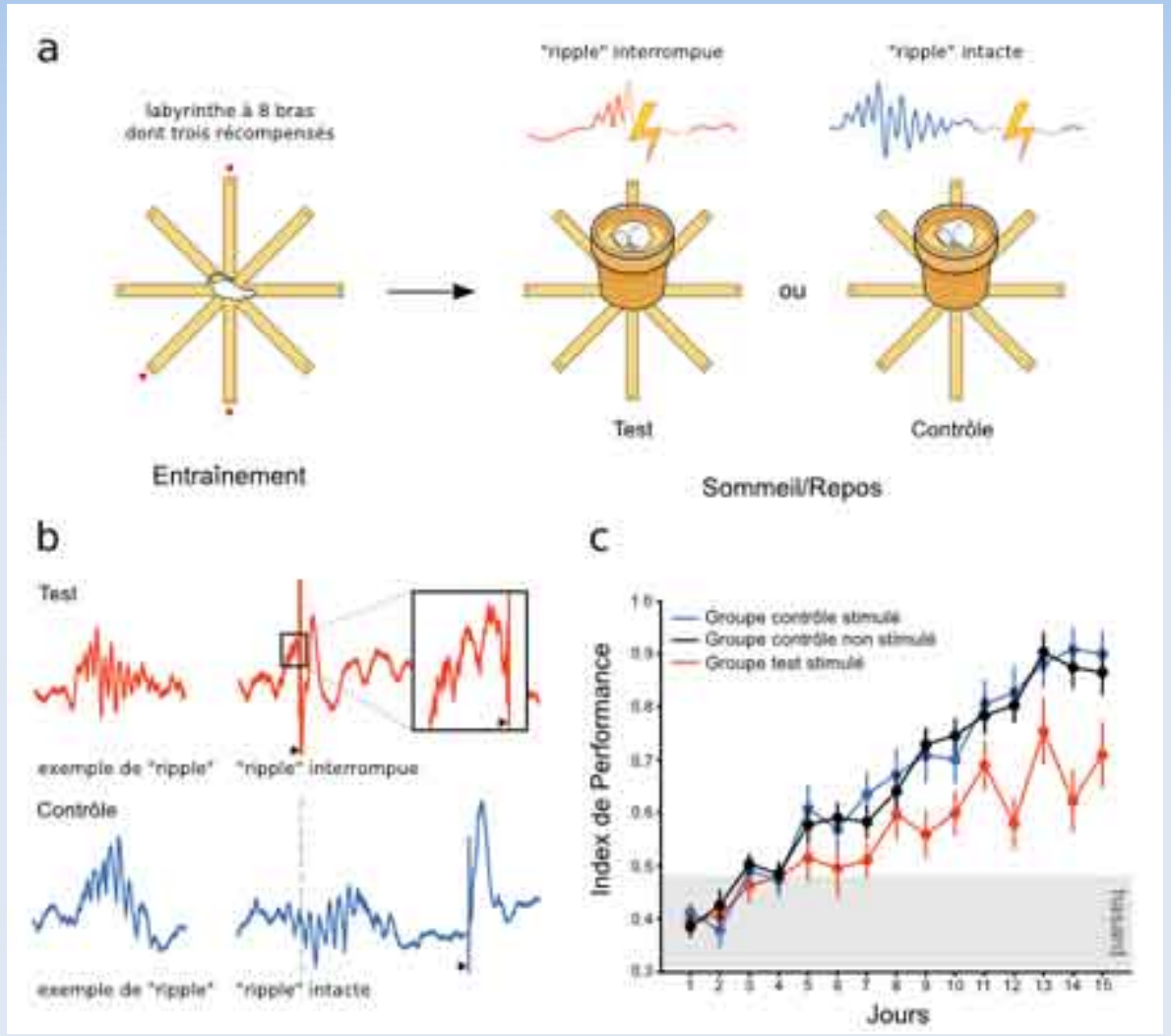
- Forward replay of hippocampal place cells during sleep (sequence is compressed 7 times) (Euston et al., 2007, Science)

Sharp-Wave Ripple (SWR) events

- "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.



Selective suppression of SWRs impairs spatial memory



- Girardeau G, Benchenane K, Wiener SI, Buzsáki G, Zugaro MB (2009) Nat Neurosci.

Hippocampal place cells

SUMMARY OF NEUROSCIENCE DATA

Replay their sequential activity during sleep (Foster & Wilson, 2006; Euston et al., 2007; Gupta et al., 2010)

Performance is impaired if this replay is disrupted (Girardeau, Benchenane et al. 2012; Jadhav et al. 2012)

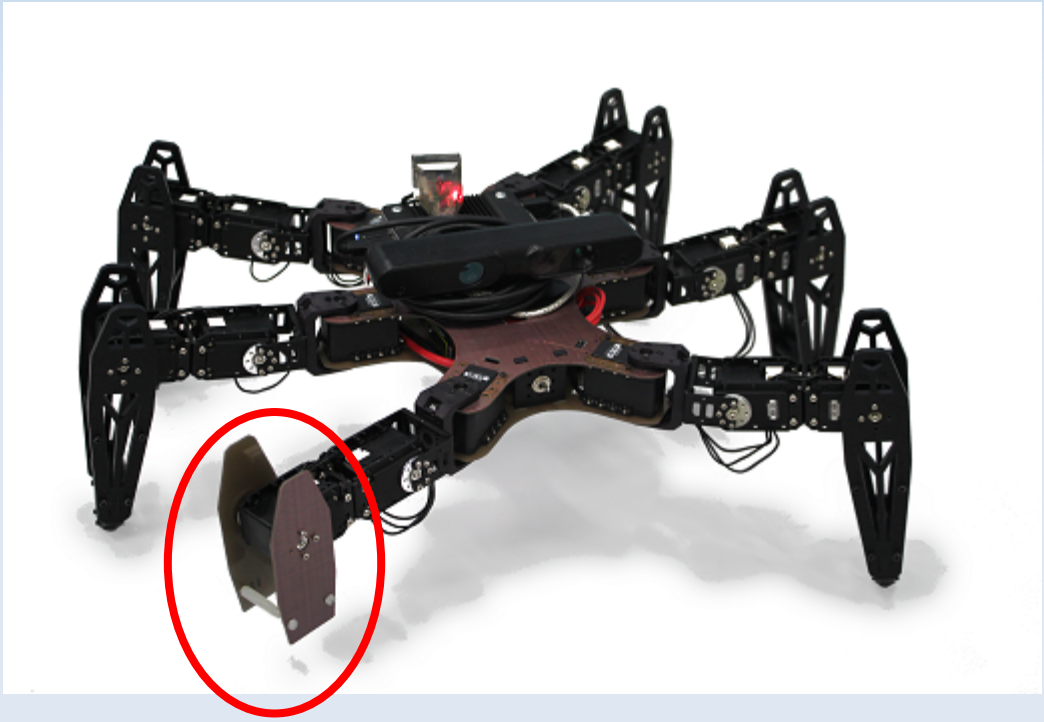
Only task-related replay in PFC (Peyrache et al., 2009)

Hippocampus may contribute to model-based navigation strategies, striatum to model-free navigation strategies (Khamassi & Humphries, 2012)

Applications to robot off-line learning

Work of Jean-Baptiste Mouret et al. @ ISIR

How to recover from damage without needing to identify the damage?

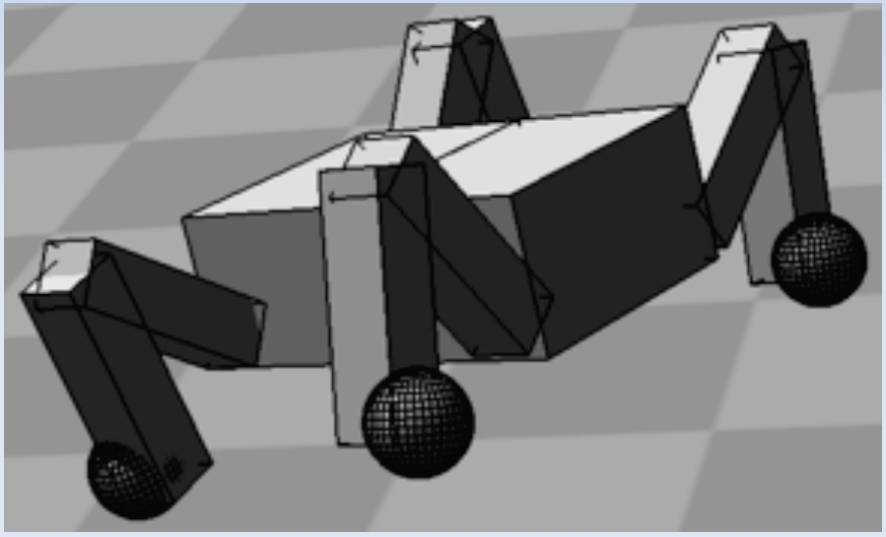


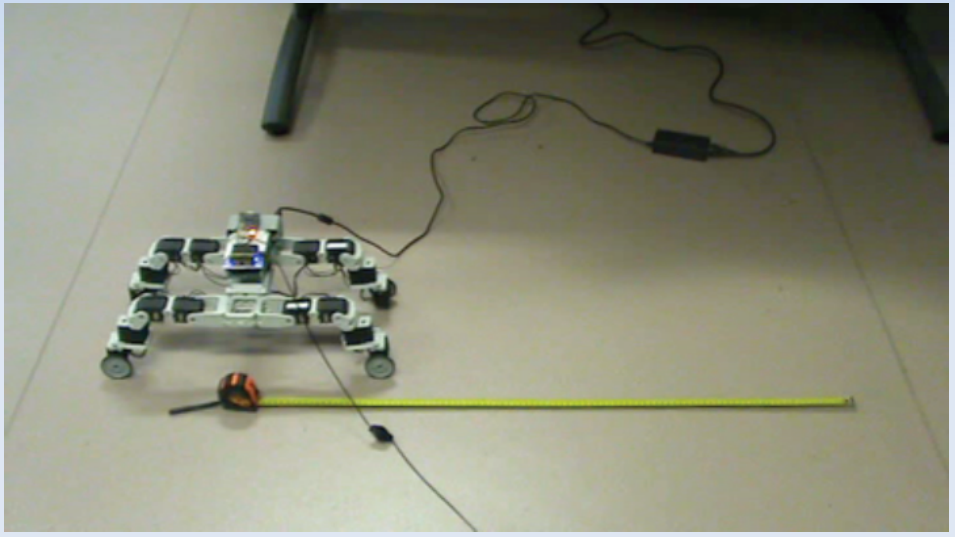
Applications to robot off-line learning

Work of Jean-Baptiste Mouret et al. @ ISIR

The reality gap

Self-model vs reality: how to use a simulator?





Solution: Learn a transferability function (how well does the simulation match reality?) with SVM or neural networks.

Idea: the damage is a large reality gap.

Koos, Mouret & Doncieux. IEEE Trans Evolutionary Comput 2012

Applications to robot off-line learning

Work of Jean-Baptiste Mouret et al. @ ISIR

Experiments

Koos, Cully & Mouret. Int J Robot Res 2013

OUTLINE

1. Intro
2. Reinforcement Learning model
   - Algorithm
   - Dopamine activity
3. Continuous RL
   - Robot navigation
   - Neuro-inspired models
4. PFC & off-line learning
   - Indirect reinforcement learning
   - Replay during sleep
5. Meta-Learning
   - Principle
   - Neuronal recordings
   - Humanoid Robot interaction

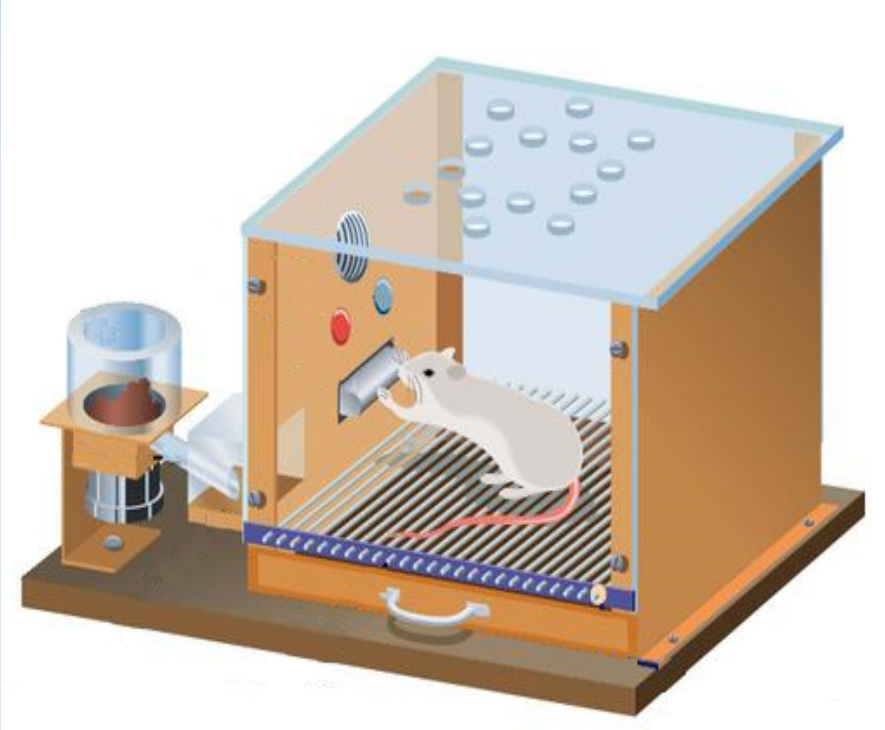
META-LEARNING

(regulation of decision-making)

1. Dual-system RL coordination
2. Online parameters tuning

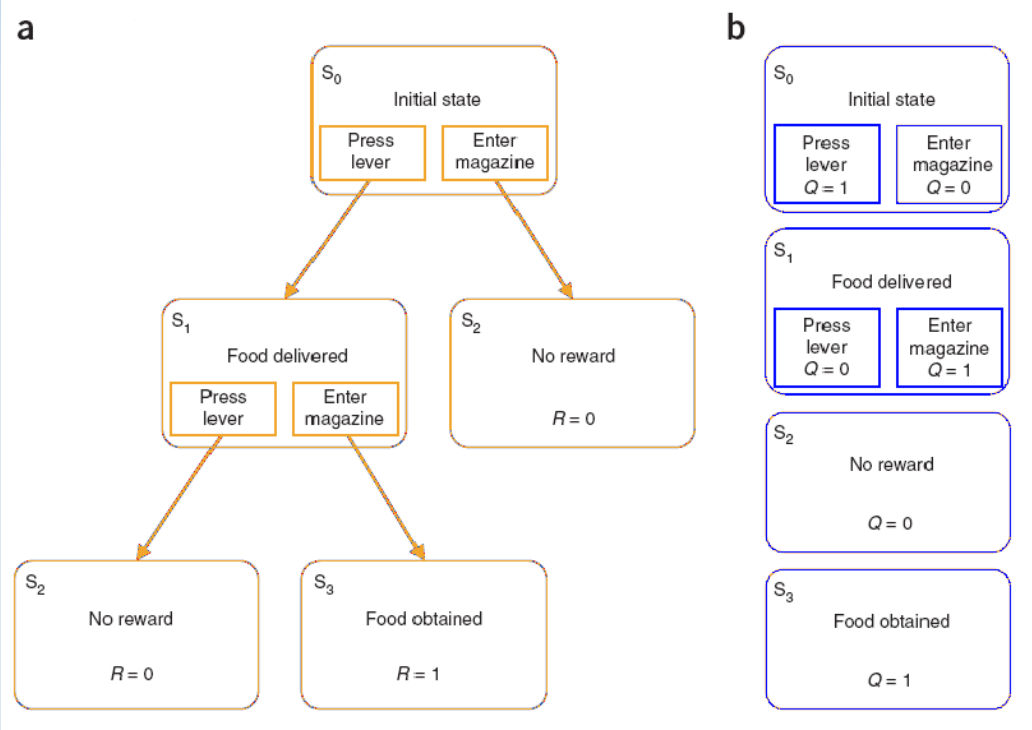
Multiple decision systems

Skinner box (instrumental conditioning)



Model-based system

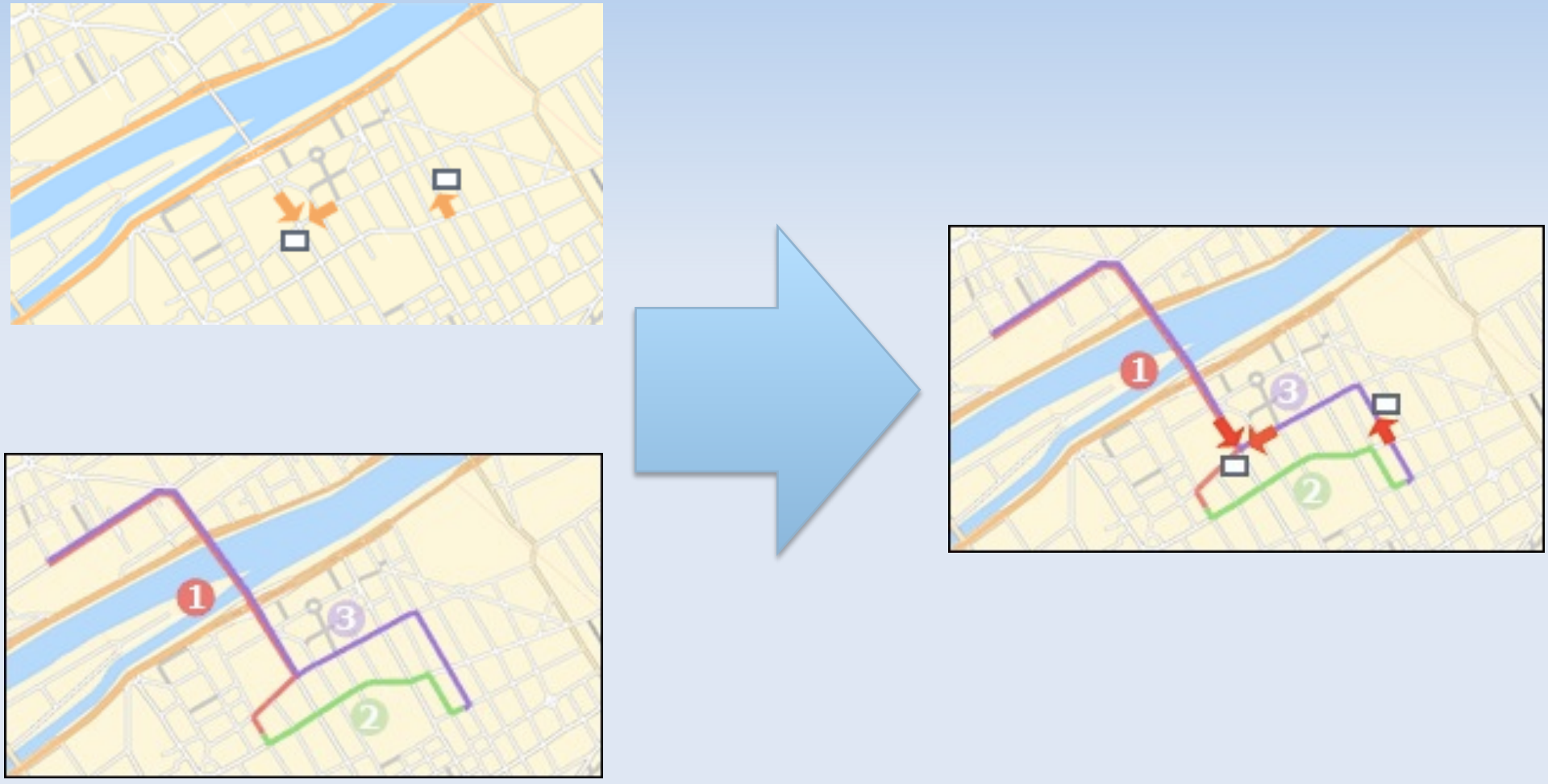
Model-free sys.



(Daw Niv Dayan 2005, Nat Neurosci)

Behavior is initially model-based and becomes model-free (habitual) with overtraining.

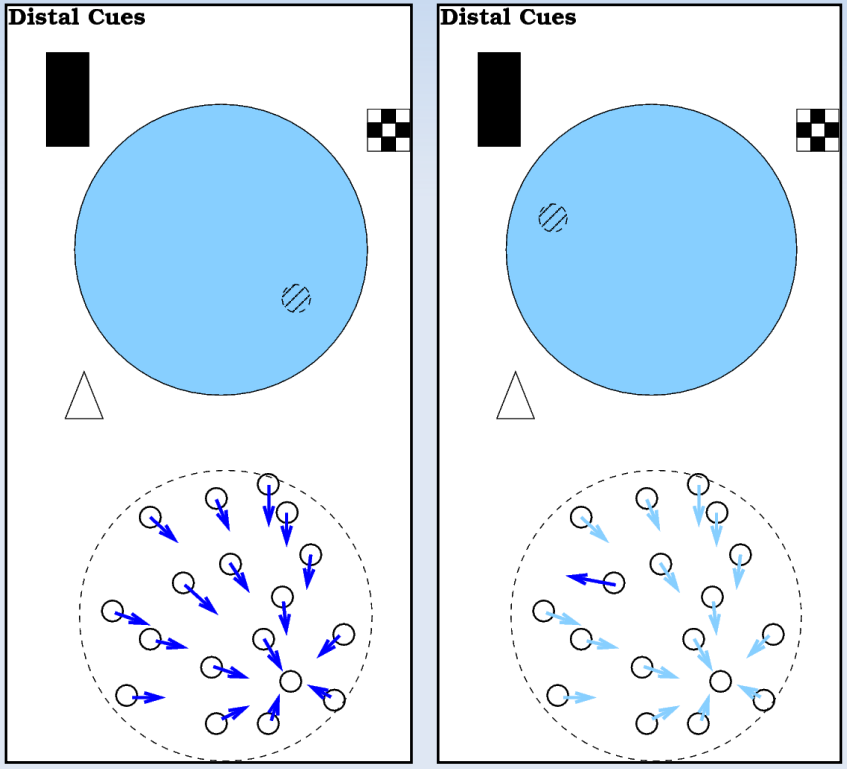
Progressive shift from model-based navigation to model-free navigation



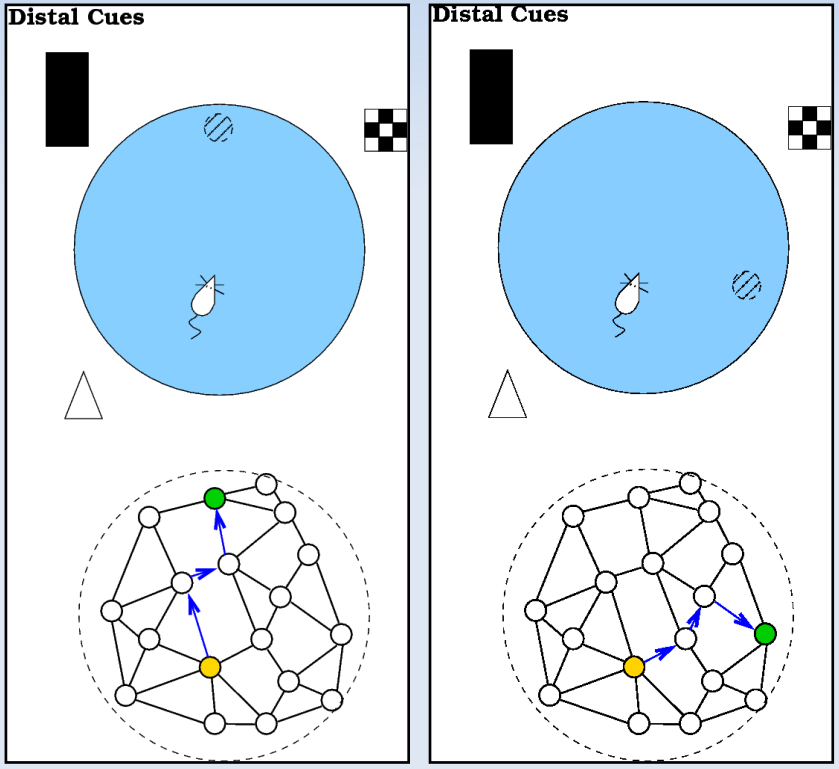
Khamassi & Humphries (2012) Frontiers in Behavioral Neuroscience

Model-based and model-free navigation strategies

Model-free navigation



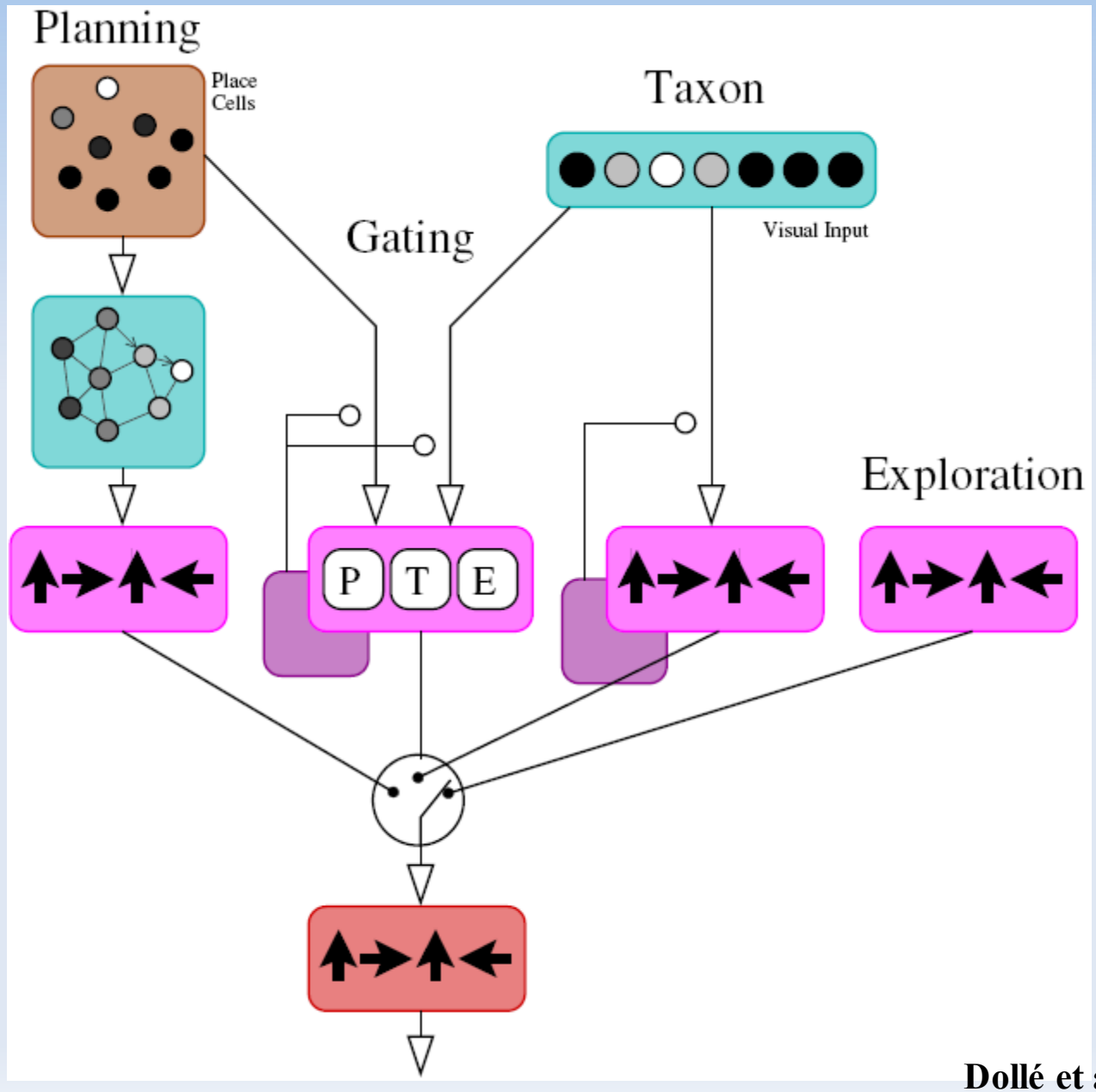
Model-based navigation



MULTIPLE DECISION SYSTEMS IN A NAVIGATION MODEL

Model-based system

(hippocampal place cells)



Model-free system

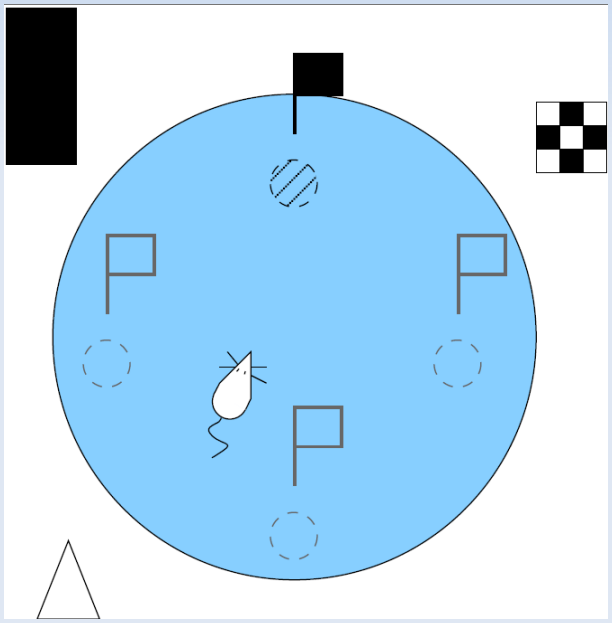
(basal ganglia)

Work by Laurent Dollé:

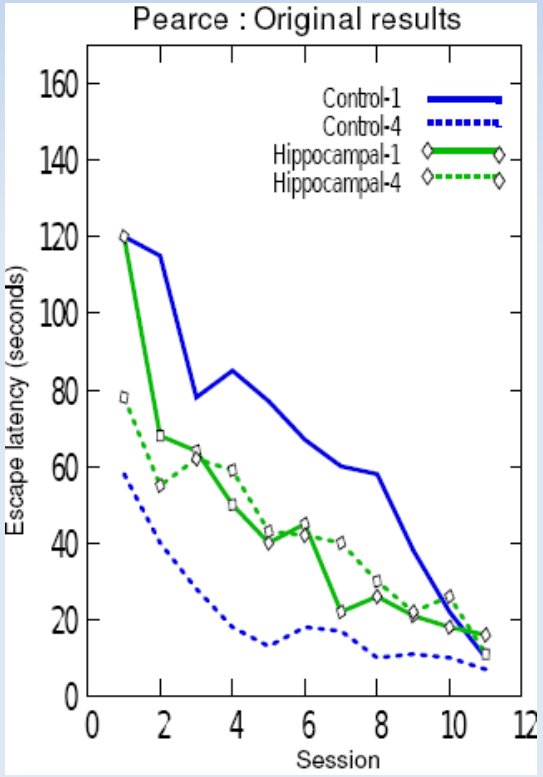
Dollé et al., 2008, 2010, submitted

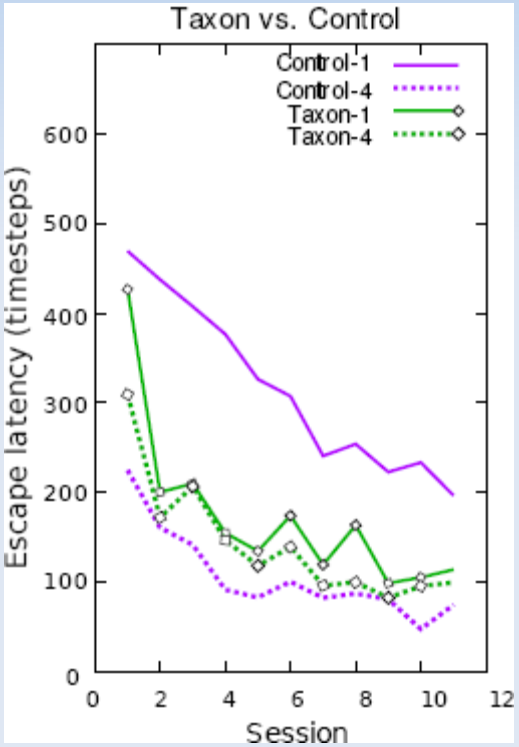
MULTIPLE NAVIGATION STRATEGIES IN A TD-LEARNING MODEL

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



Task of Pearce et al., 1998

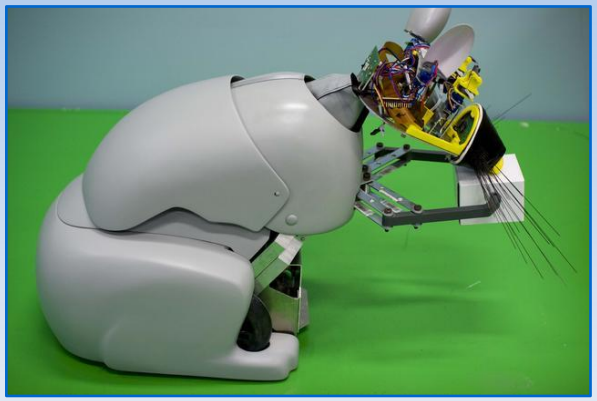




Model:

Dollé et al., 2010

PSIKHARPAX ROBOT



Work by:

Ken Caluwaerts (2010)

Steve N'Guyen (2010)

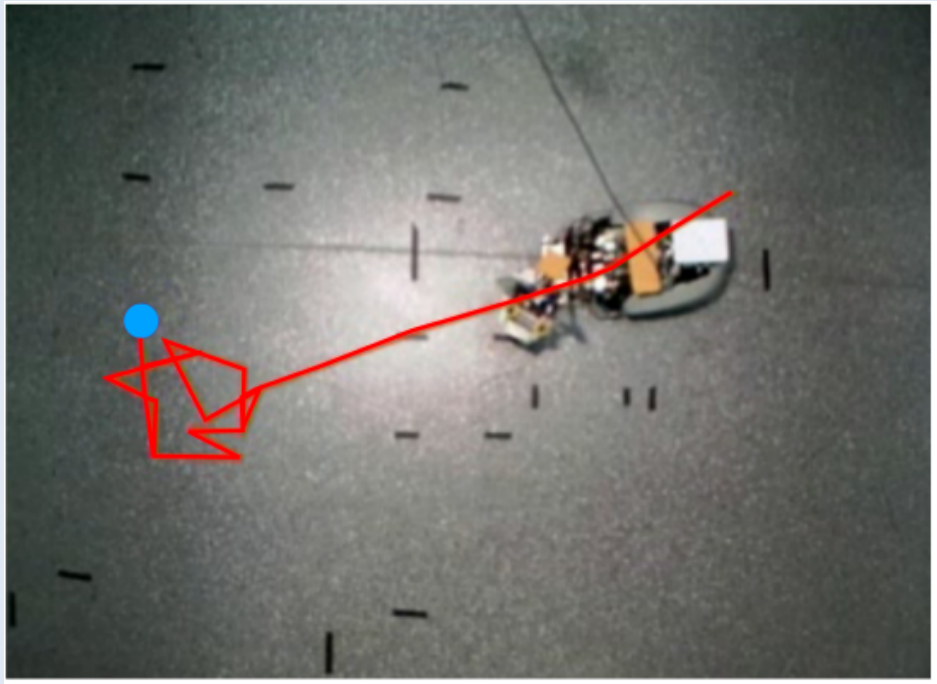
Mariacarla Staffa (2011)

Antoine Favre-Félix (2011)

Caluwaerts et al. (2012) Biomimetics & Bioinspiration

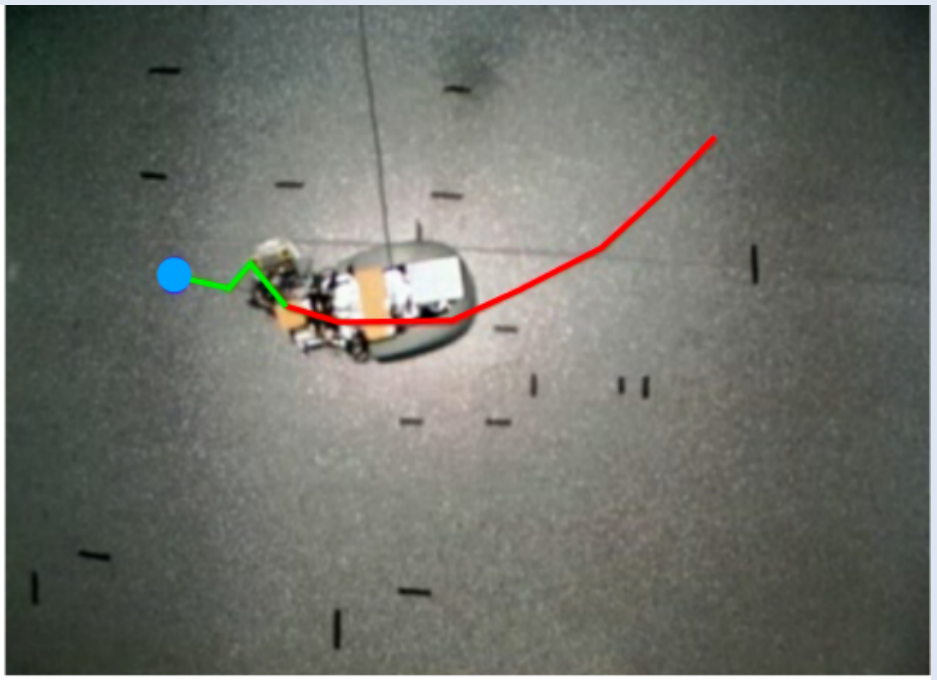
PSIKHARPAX ROBOT

Planning strategy only



(*a*)

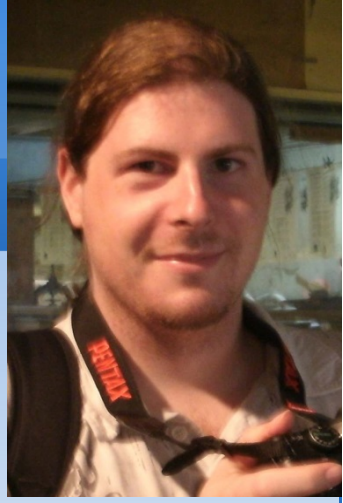
Planning strategy + Taxon strategy

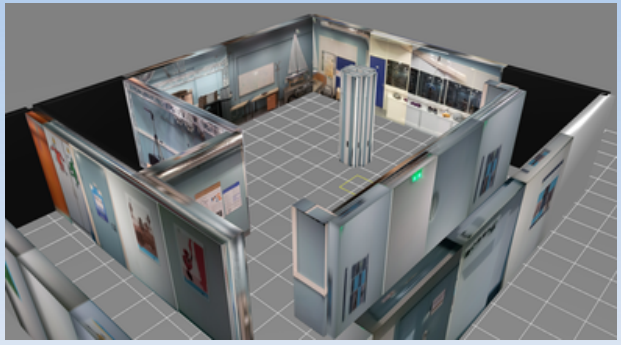


(*b*)

Caluwaerts et al. (2012) Biomimetics & Bioinspiration

CURRENT APPLICATIONS TO THE PR2 ROBOT

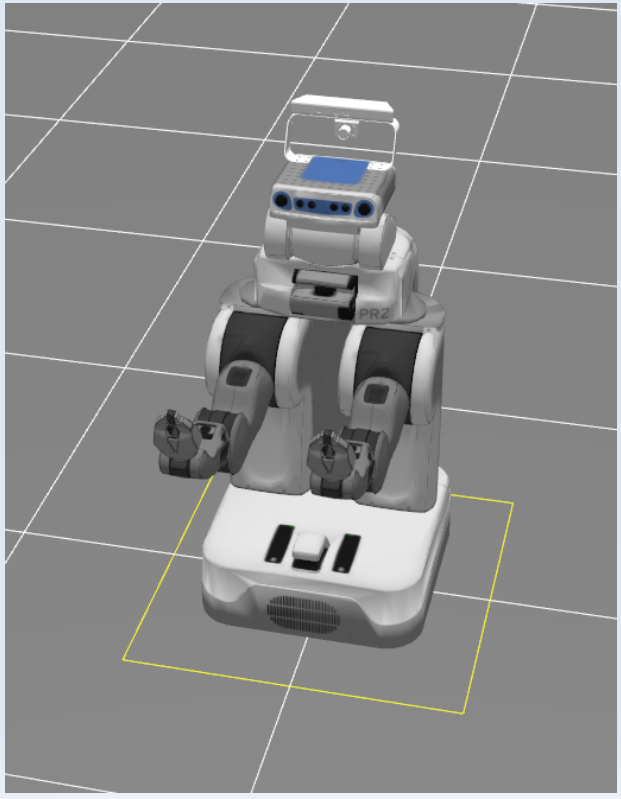


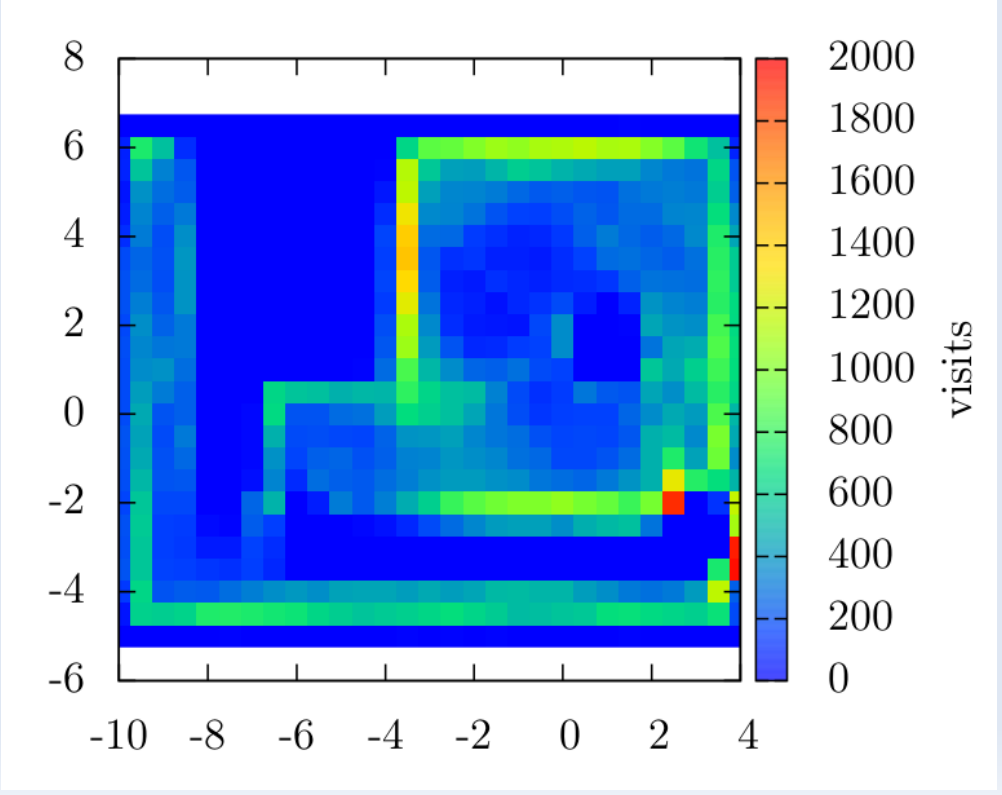


Travaux de :

Erwan Renaudo

Omar Islas Ramirez





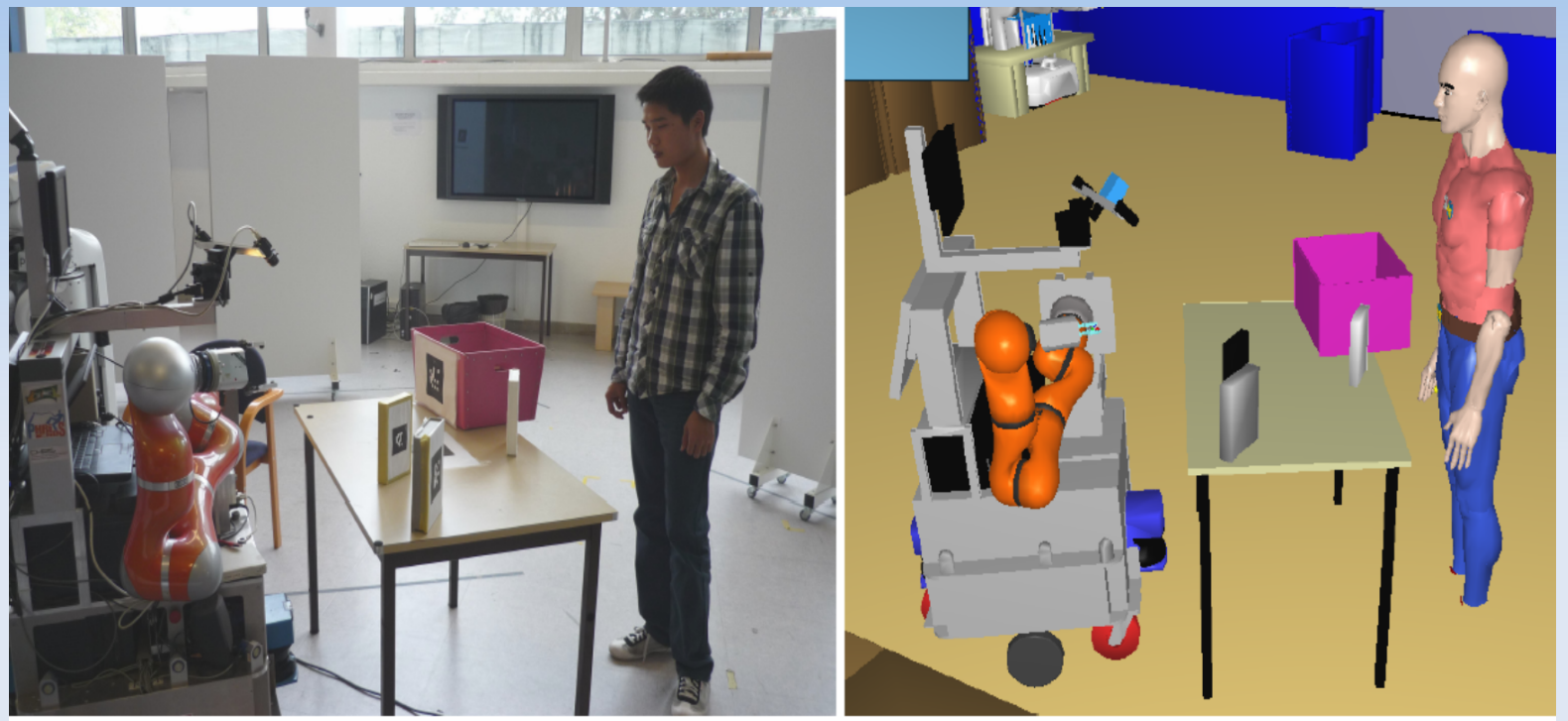
CURRENT APPLICATIONS TO HUMAN-ROBOT INTERACTION

Travaux de :

Erwan Renaudo

Collaboration :

Alami et al (LAAS)

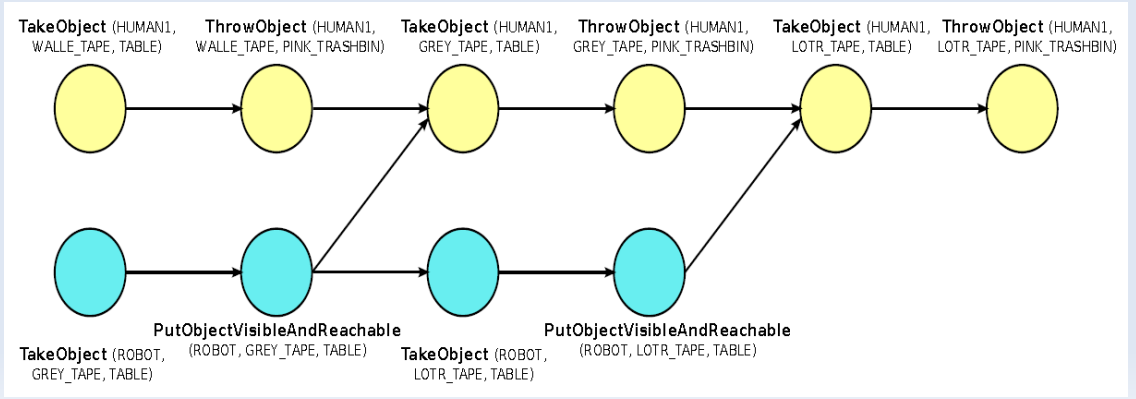


(a) Initial state (b) 3d model view of initial state

Task: Clean the table

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

Goal: Autonomous learning by the robot



META-LEARNING

(regulation of decision-making)

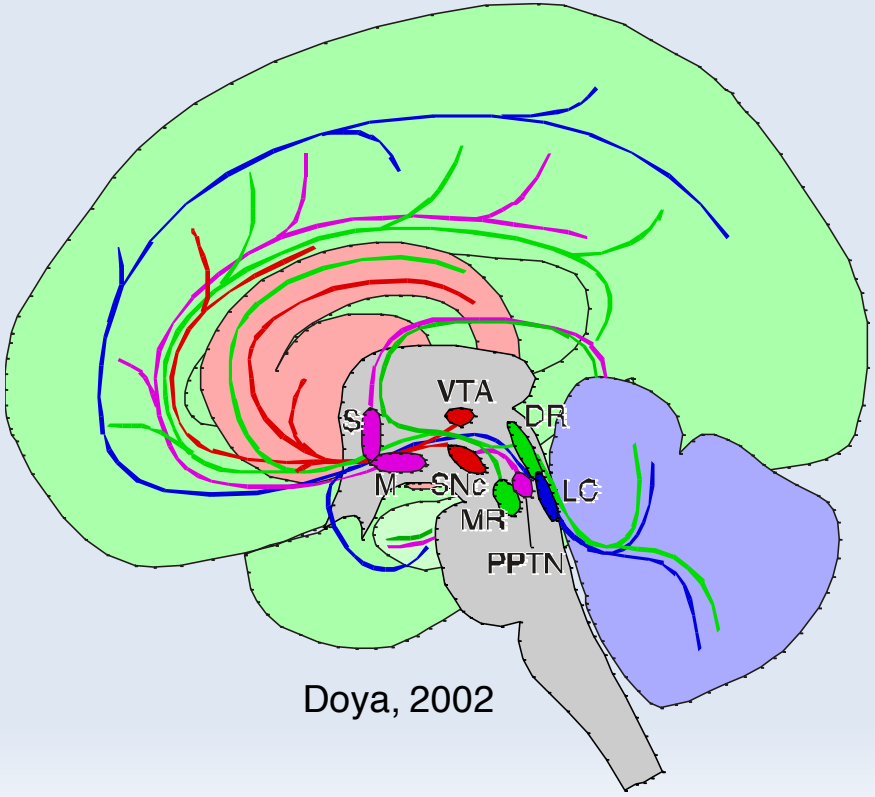
1. Dual-system RL coordination
2. Online parameters tuning

REINFORCEMENT LEARNING & META-LEARNING FRAMEWORK

$$Q(s,a) \leftarrow Q(s,a) + \alpha \,.\, \delta$$ **Action values update**

$$\delta = r + \gamma \,.\, \max[Q(s',a')] - Q(s,a)$$ **Reinforcement signal**

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



Doya, 2002

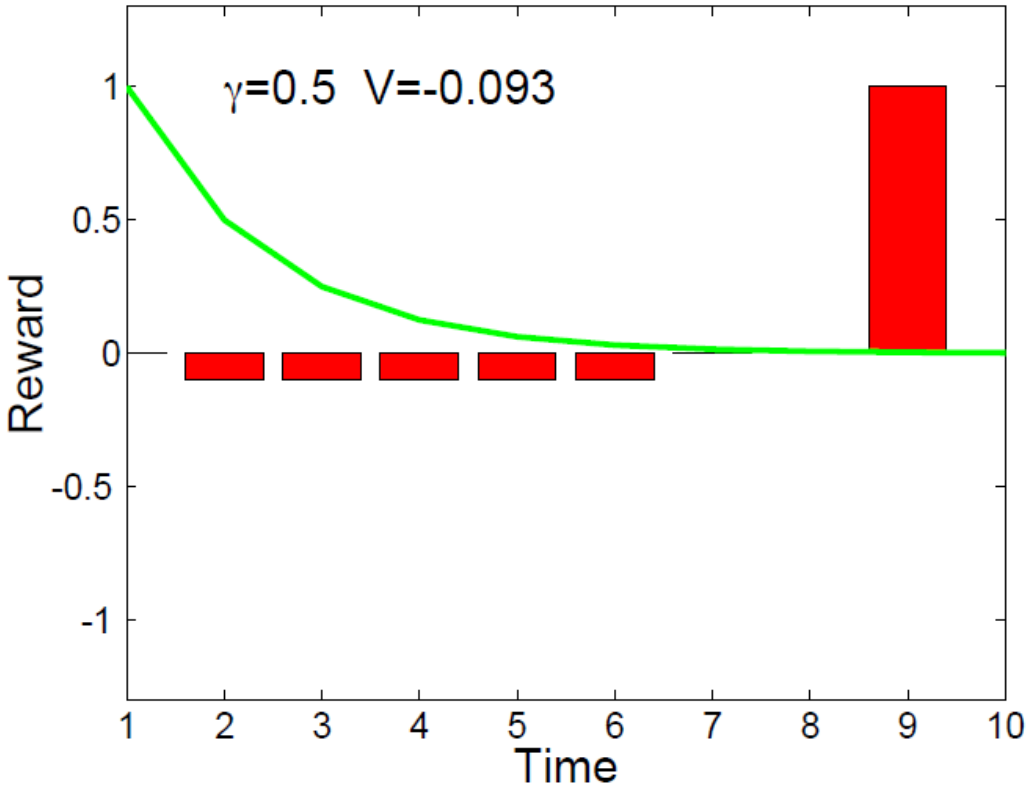
Dopamine: TD error δ

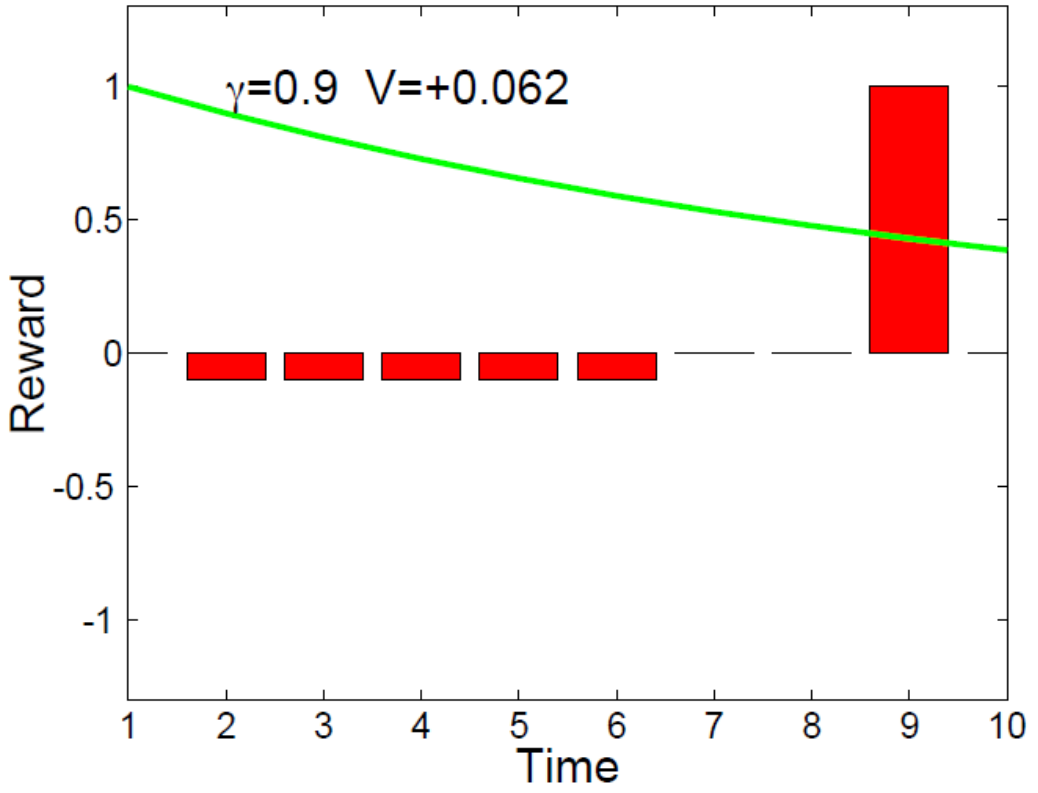
Acetylcholine: learning rate α

Noradrenaline: exploration β

Serotonin: temporal discount γ

META-LEARNING

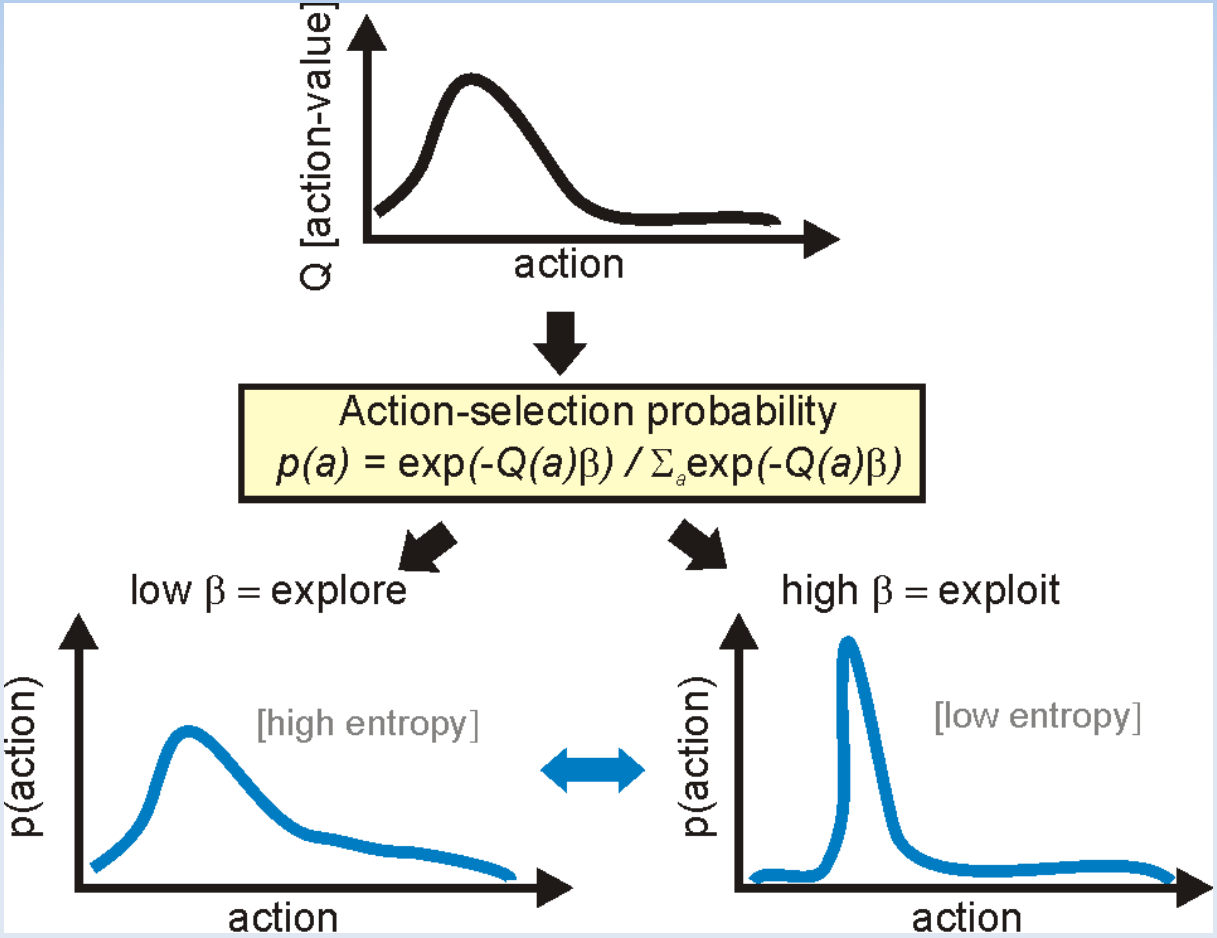




Doya, 2002

Effect of γ on expected reward value

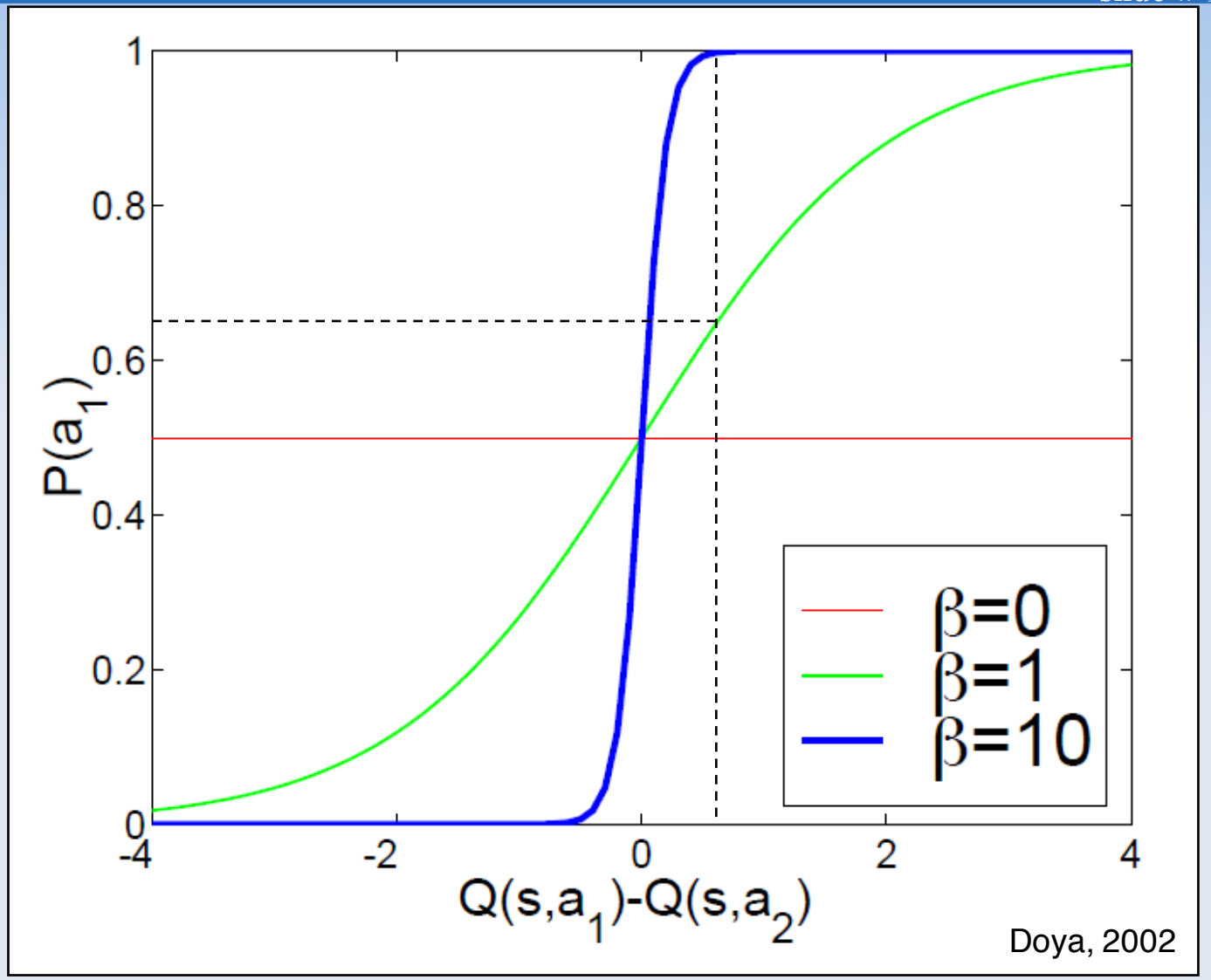
META-LEARNING



- The exploration-exploitation trade-off: necessary for learning; but impacts on action selection.

META-LEARNING

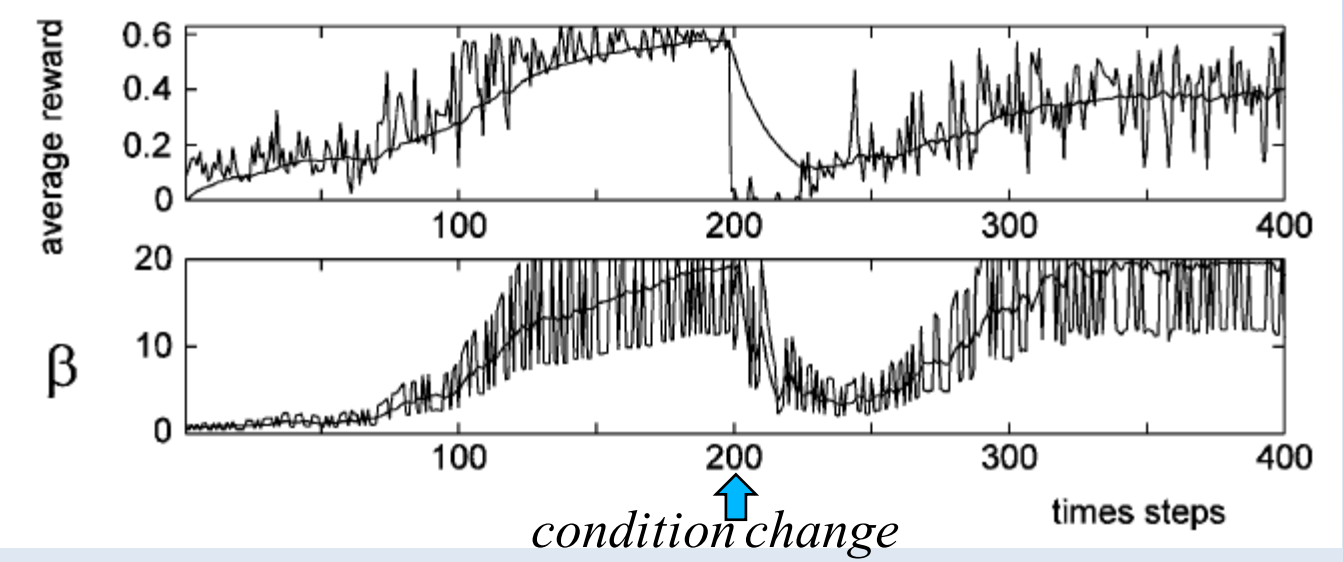
Effect of β on exploration



Boltzmann softmax equation: $P(a) = \frac{\exp(\boldsymbol{\beta} \cdot Q(s,a))}{\sum_b \exp(\boldsymbol{\beta} \cdot Q(s,b))}$

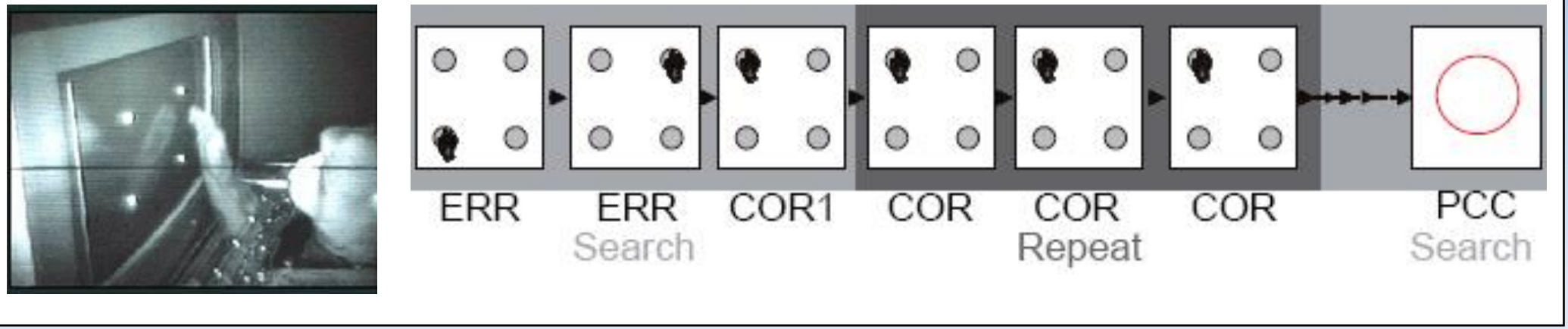
META-LEARNING

- Meta-learning methods propose to tune RL parameters as a function of average reward and uncertainty (Schweighofer & Doya, 2003).



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

TASK

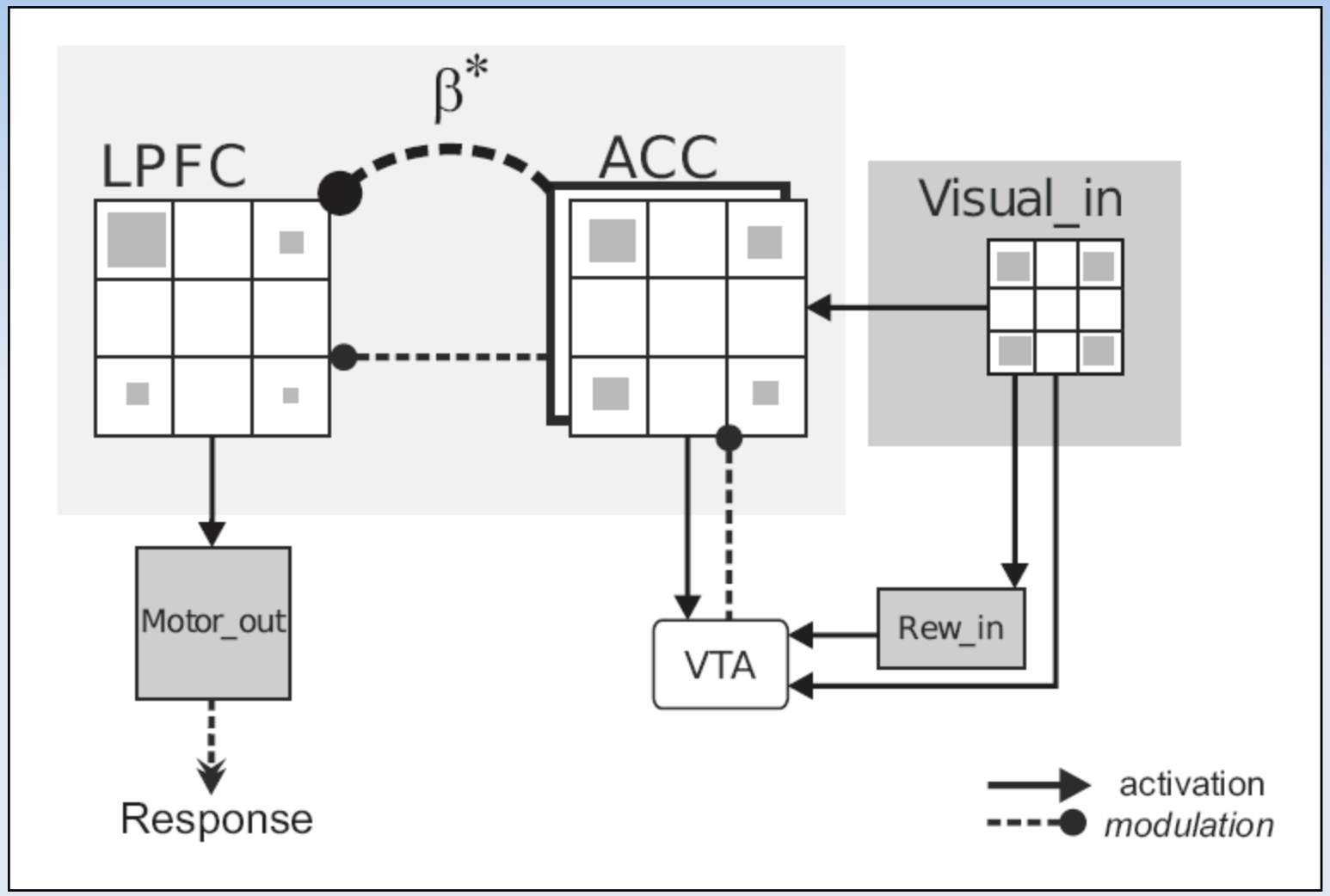


Question: How did the monkeys learn to re-explore after each presentation of the PCC signal?

Hypothesis: By trial-and-error during pretraining.

Khamassi et al. (2011) Front in Neurorobotics; Khamassi et al. (2013) Prog Brain Res

Computational model

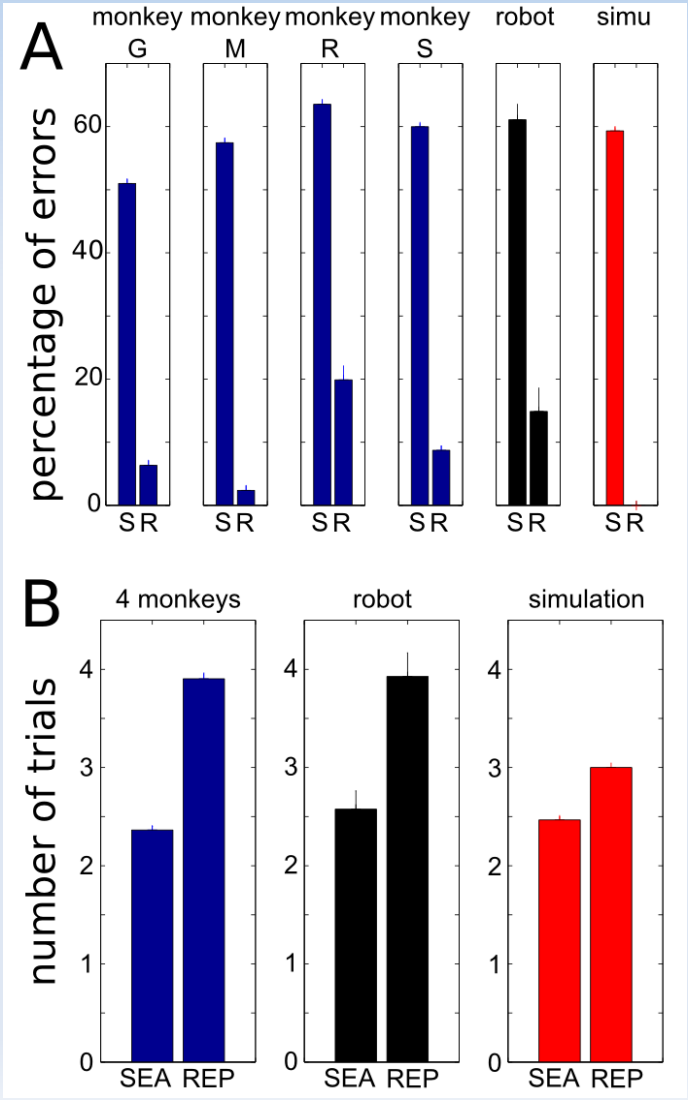


β^*: exploratory variable used to modulate β

Khamassi et al. (2011) Frontiers in Neurorobotics

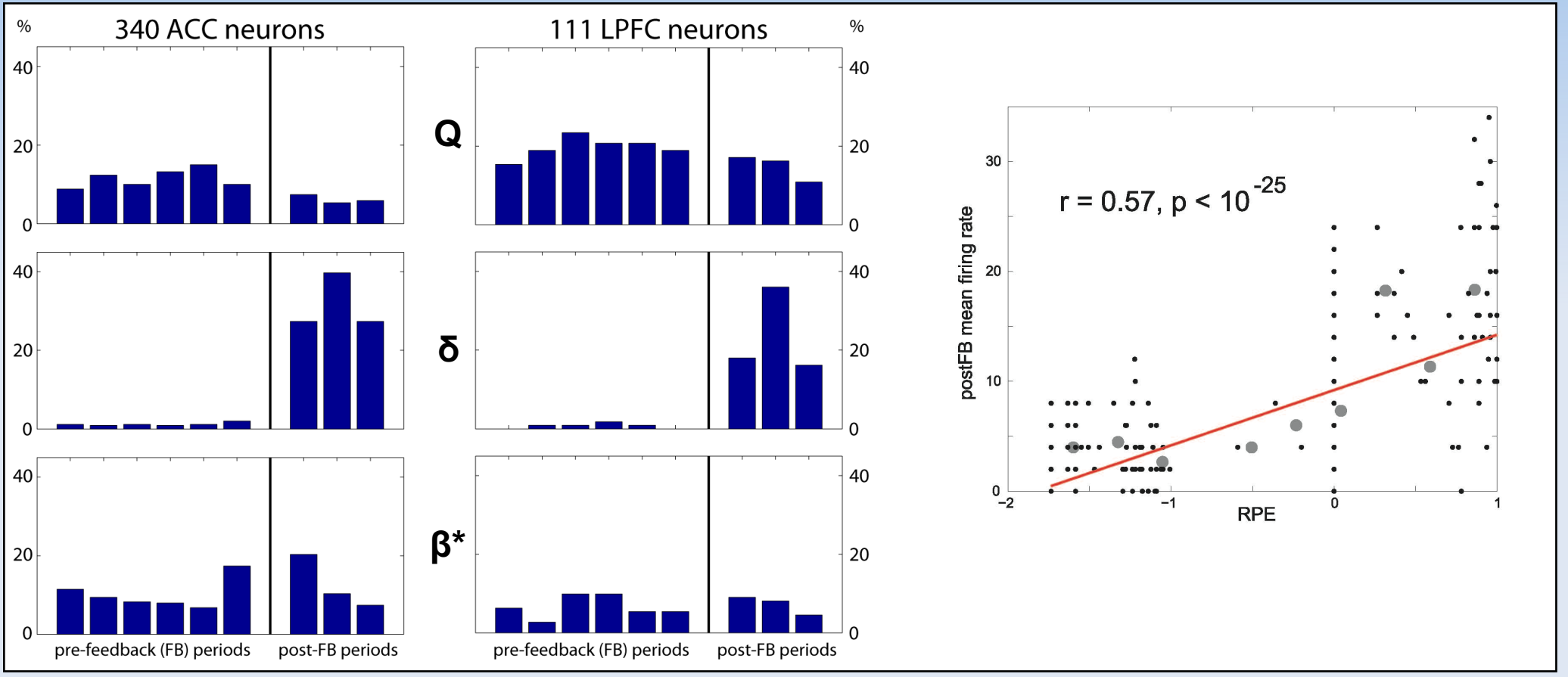
Computational model

- Reproduction of the global properties of monkey performance in the PS task.



Khamassi et al. (2011) Frontiers in Neurorobotics

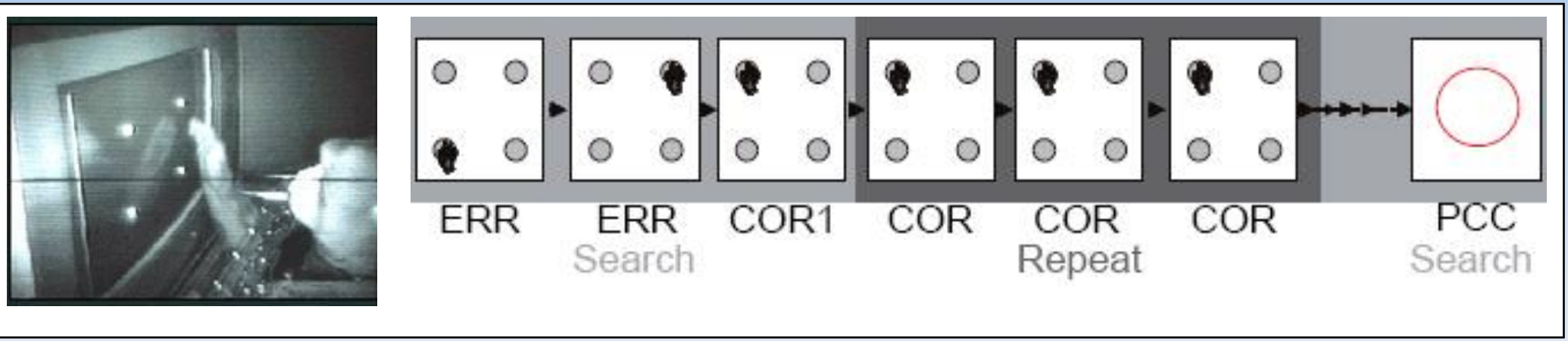
Model-based analysis
My post-doc work



Multiple regression analysis with bootstrap

Khamassi et al. (2013) Prog Brain Res; Khamassi et al. (in revision)

Meta-learning applied to Human-Robot Interaction



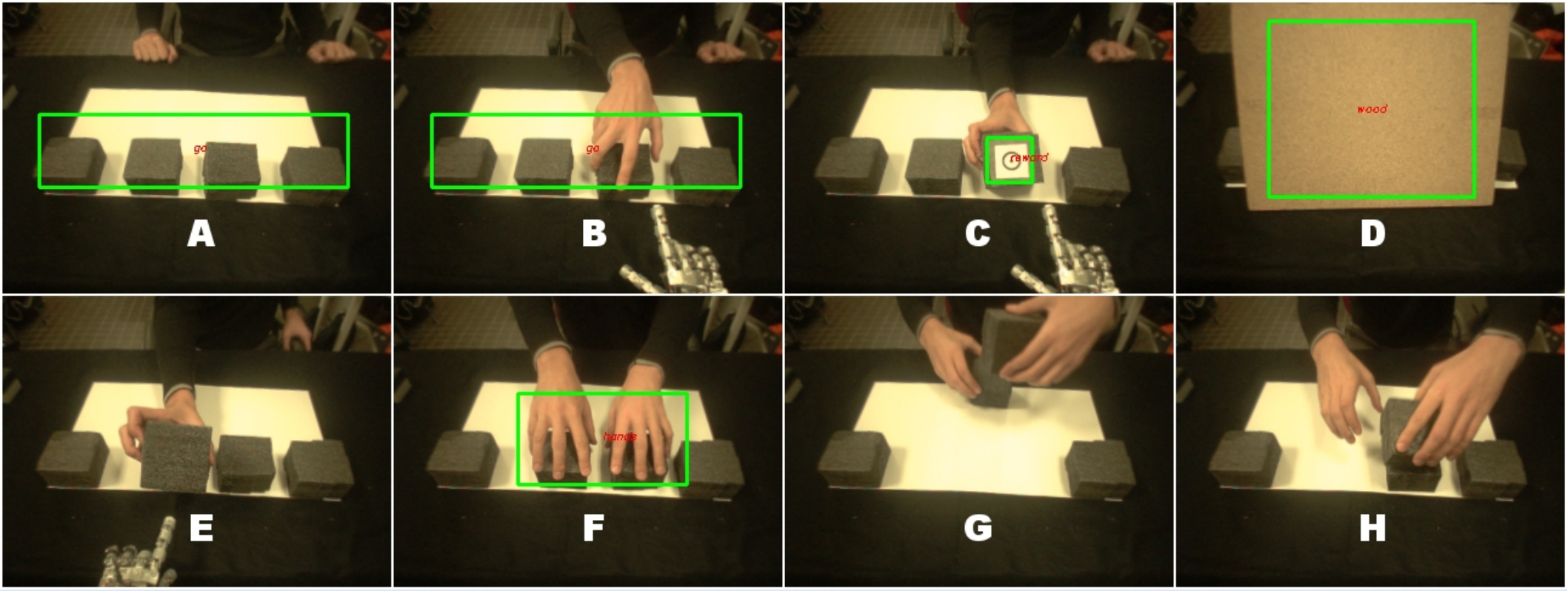
- In the previous task, monkeys and the model a priori 'know' that *PCC* means a reset of exploration rate and action values.
- Here, we want the iCub robot to learn it by itself.

Meta-learning applied to Human-Robot Interaction

Khamassi et al. (2011) Frontiers in Neurorobotics

Meta-learning applied to Human-Robot Interaction

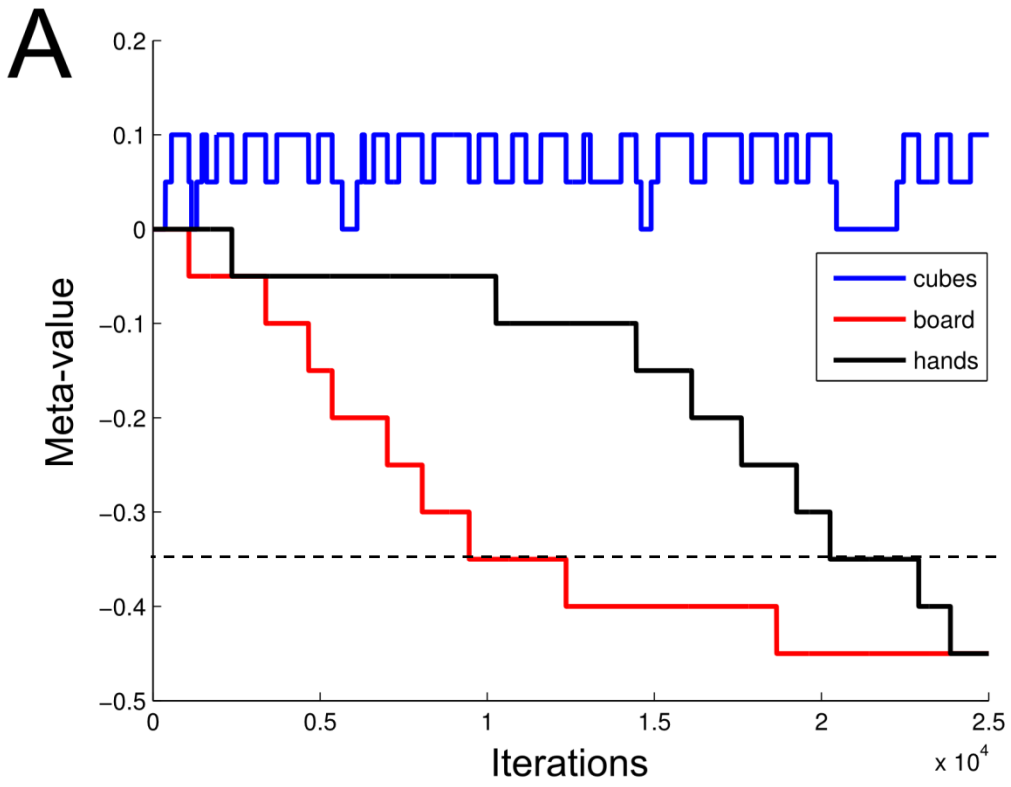
Go signal | Choice | Reward | Wooden board

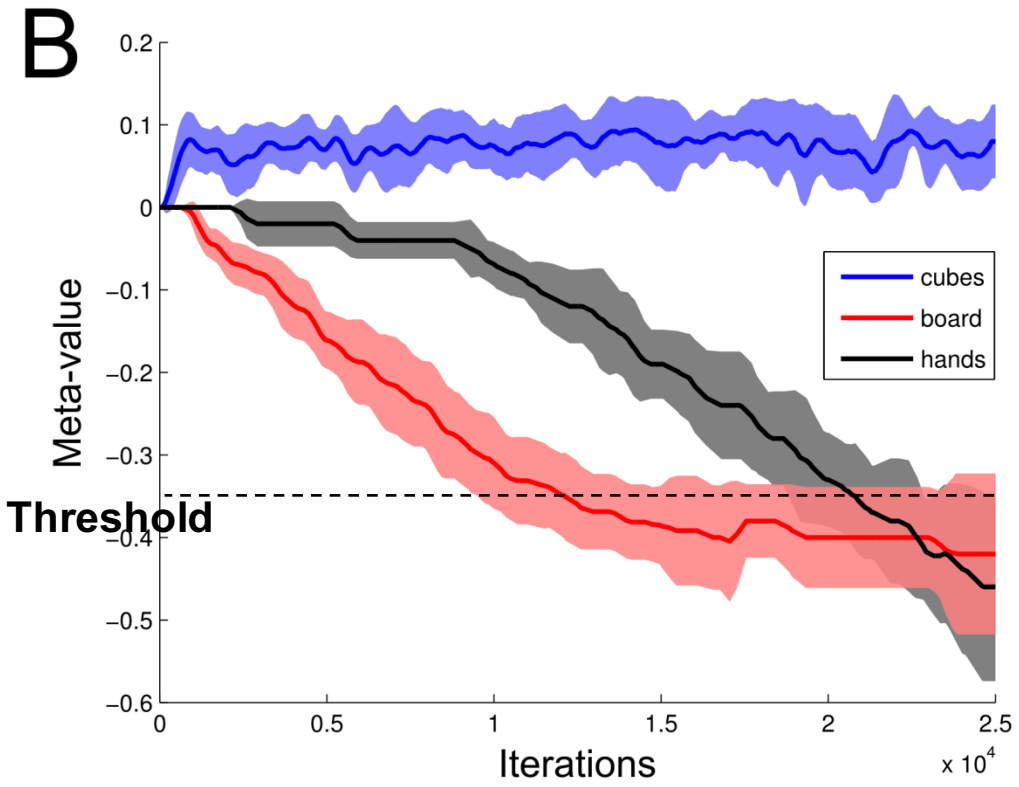


Error | Human's hands | Cheating | Cheating

Meta-learning applied to Human-Robot Interaction

meta-value(i) ← meta-value(i) + α'. Δ[averageReward]





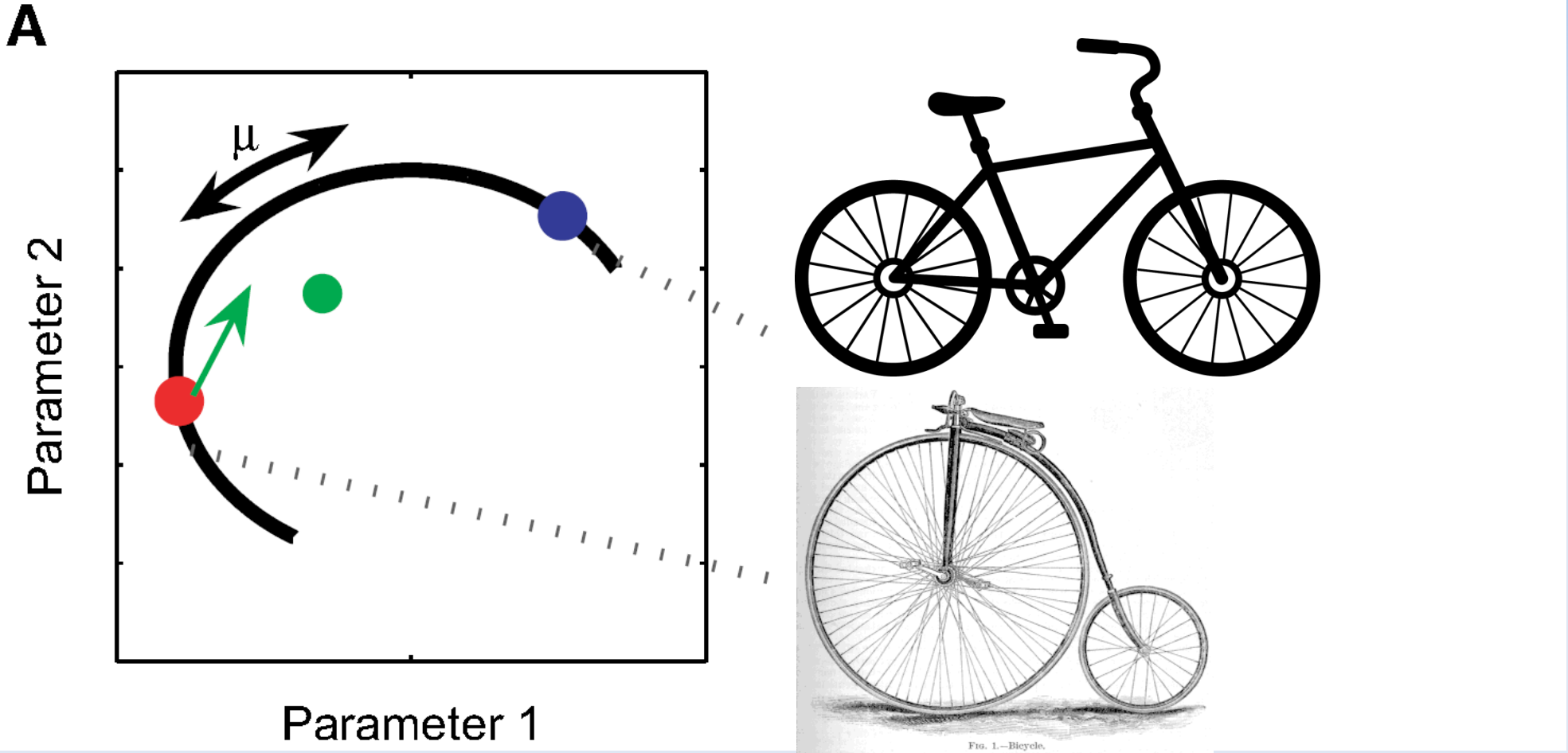
CONCLUSION OF THE ACC-LPFC META-LEARNING PART

- ACC is in an appropriate position to evaluate feedback history to modulate the exploration rate in LPFC.
- ACC-LPFC interactions could regulate exploration based on mechanisms capturable by the meta-learning framework.
- Such modulation could be subserved via noradrenaline innervation in LPFC.
- Such a pluridisciplinary approach can contribute both to a better understanding of the brain and to the design of algorithms for autonomous decision-making.

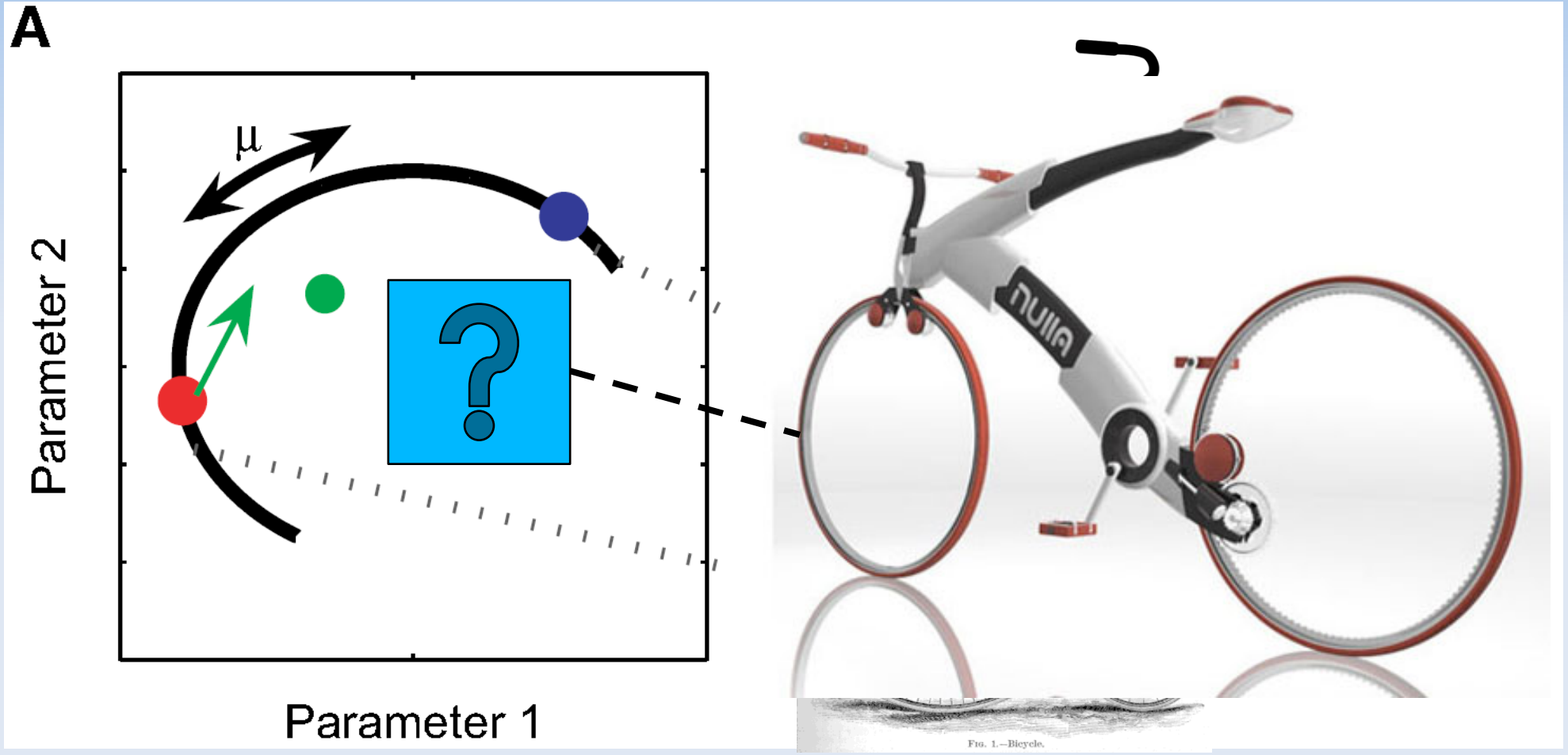
Meta-learning and motor learning

- Can meta-learning principles be useful for the integration of reinforcement learning and motor learning?

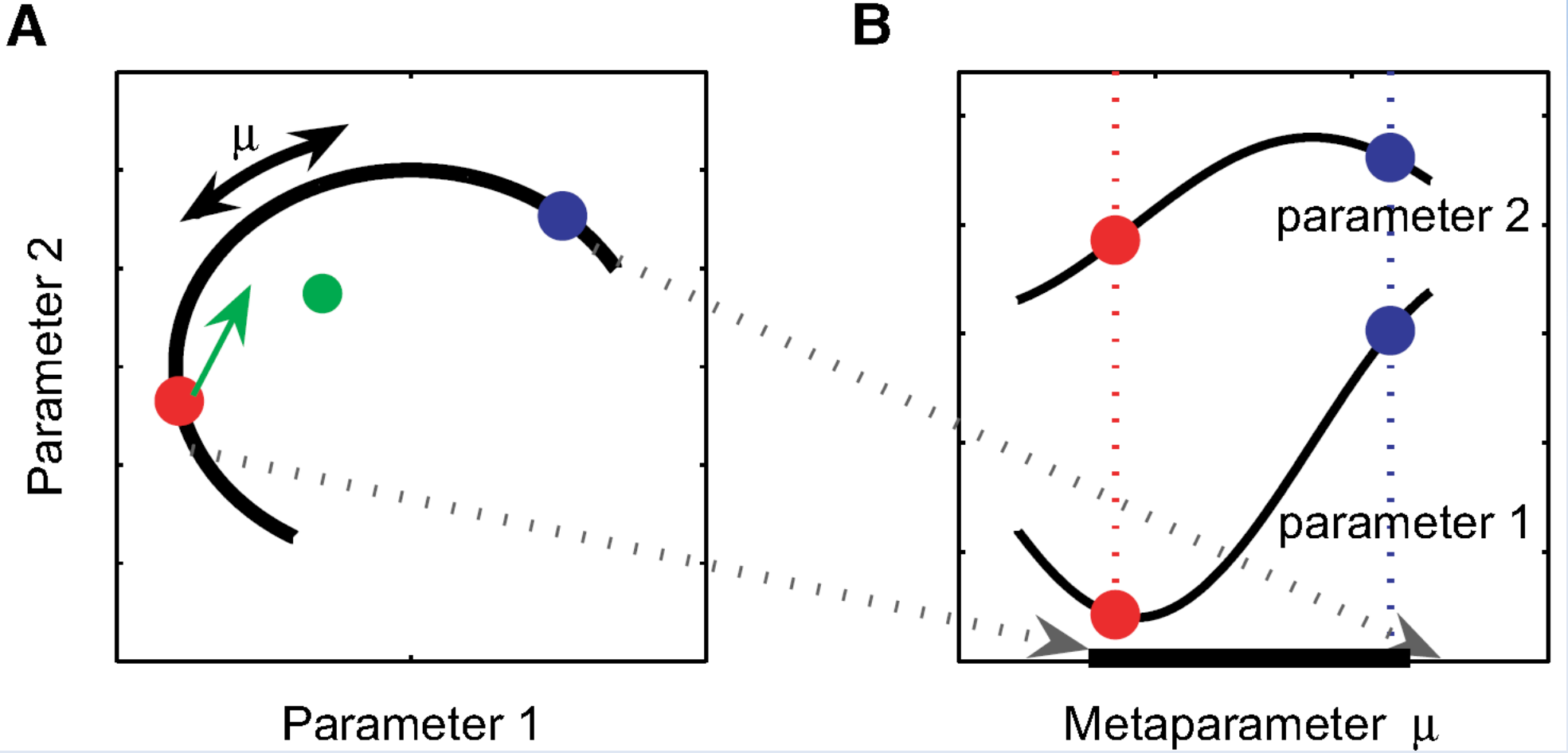
Structure learning
(Braun Aertsen Wolpert Mehring 2009)



Structure learning
(Braun Aertsen Wolpert Mehring 2009)

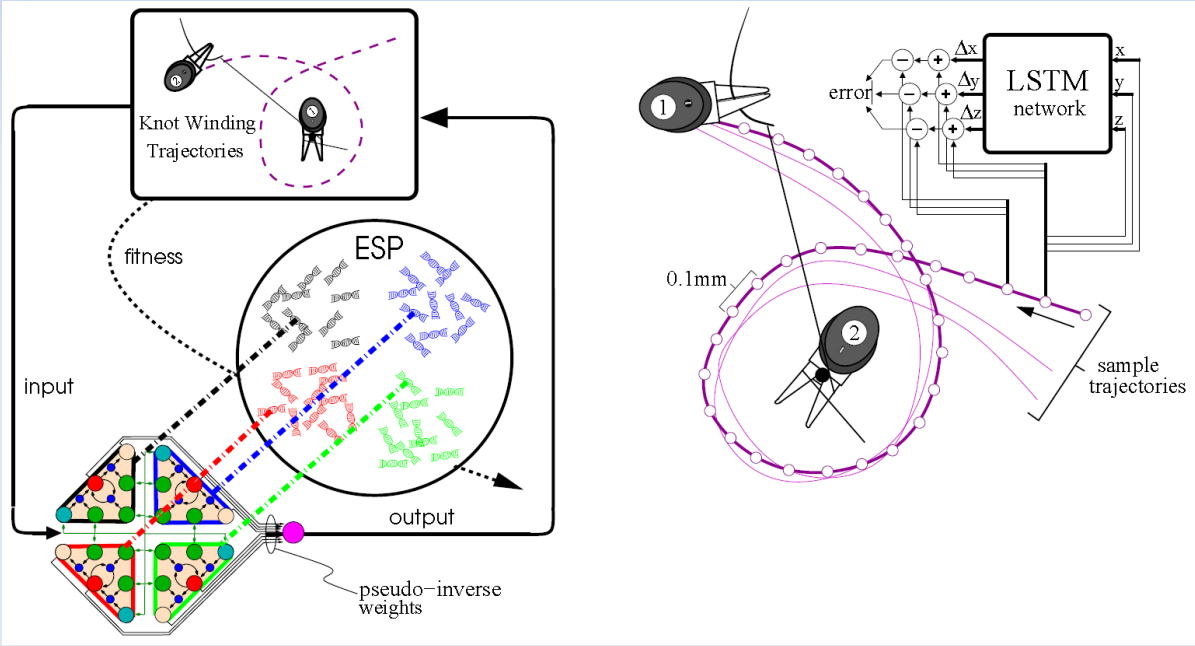


Structure learning
(Braun Aertsen Wolpert Mehring 2009)



Schmidhuber on meta-learning (1)

- Recurrent neural-networks applied to Robotics



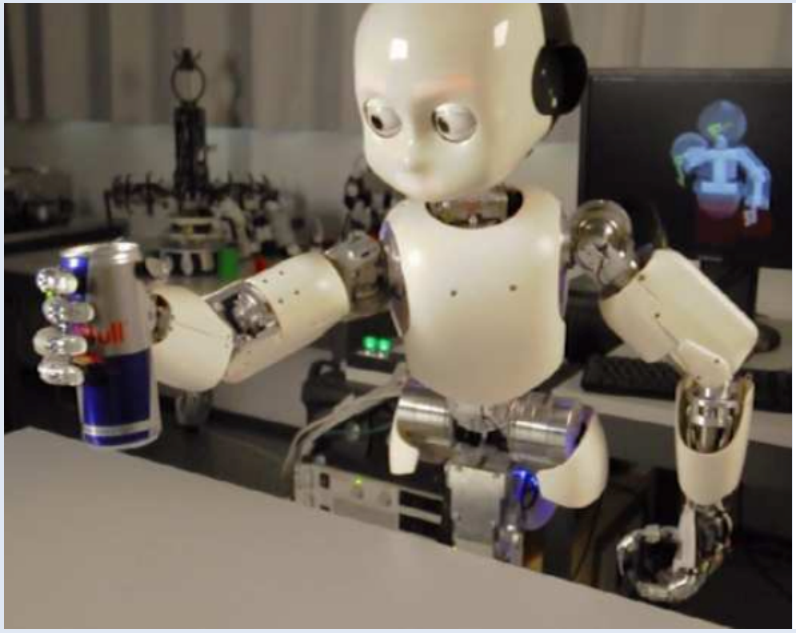
Mayer et al. (IROS 2006)

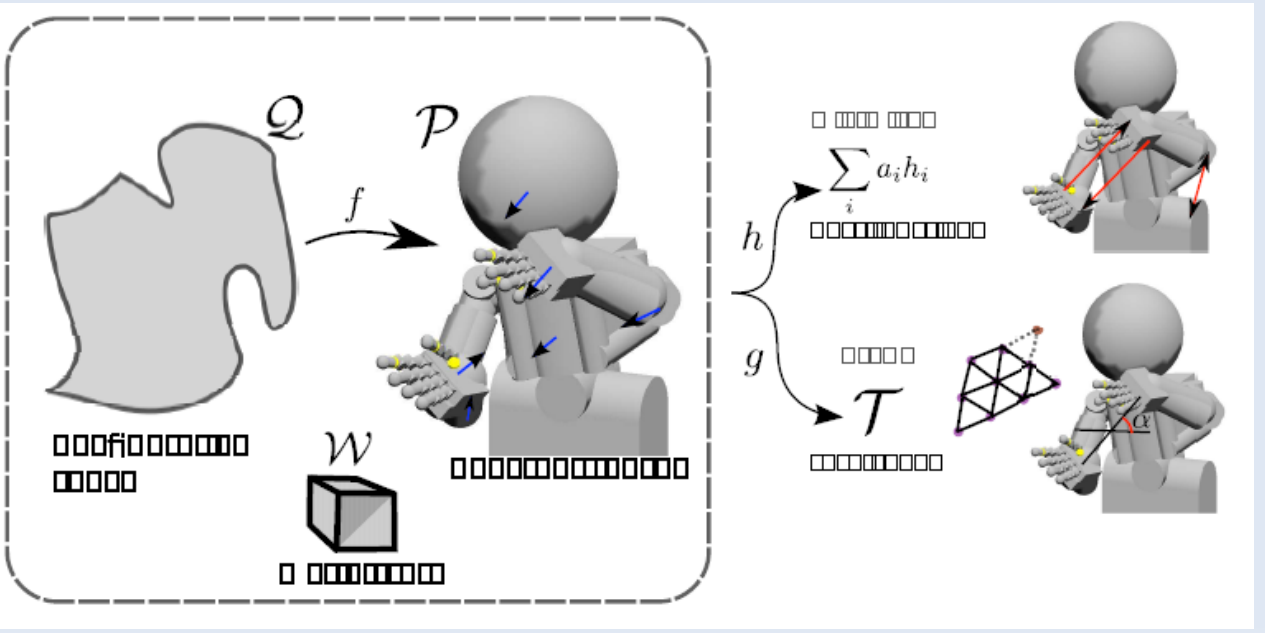
Schmidhuber on meta-learning (2)

- RL with self-modifying policies (actions that can edit the policy itself)
- Success-story criterion (time varying set V of past checkpoints that led to long-term reward accelerations)

Schmidhuber on motor learning

- Learning maps of task-relevant motor behaviors under specified constraints (e.g. maintain hands parallel ; do not touch box nor table ; …)
- How can these primitive constrained motor behaviors be used by decision system and high-level goal-directed learning?





Stollenga et al. (IROS 2013)

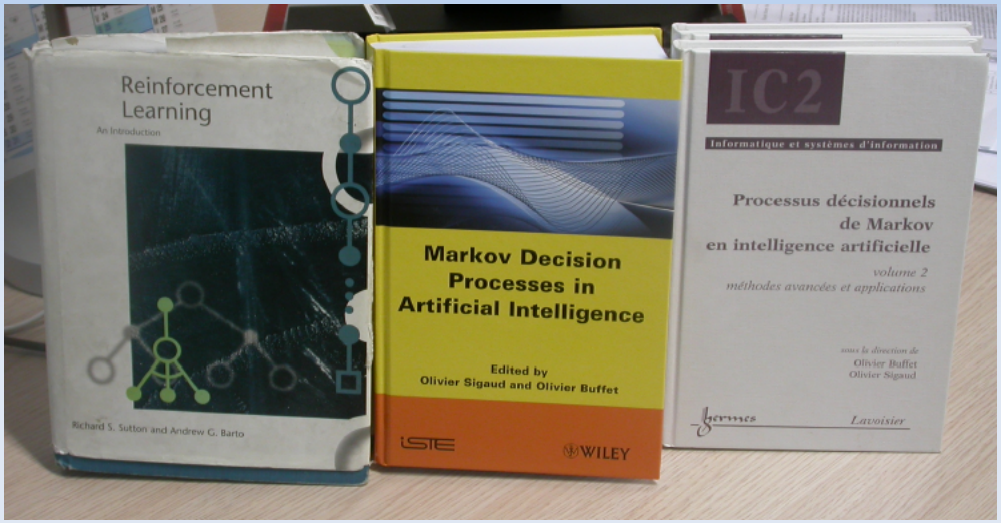
SUMMARY

- Direct RL with Temporal-Difference methods:
  - Actor-Critic / SARSA / Q-learning
  - Works well for perfect discrete state/action spaces
- Indirect RL (planning, dyna-Q, off-line learning)
  - Needs to know the transition & reward functions
- Partially Observable MDP (POMDP)
  - When the Markov hypothesis is violated (perceptual aliasing, multi-agents, non stationnary environment)
- Current advancement of RL models for:
  - continuous action space (gradient descent)
  - multiple parallel decision systems.
  - meta-learning (ACC-LPFC interactions).

CONCLUSION

- The Reinforcement Learning framework provides algorithms for autonomous agents.
- It can also help explain neural activity in the brain.
- Such a pluridisciplinary approach can contribute both to a better understanding of the brain and to the design of algorithms for autonomous decision-making.

FURTHER READINGS



1. Sutton & Barto (1998) RL: An Introduction
2. Buffet & Sigaud (2008) en français
3. Sigaud & Buffet (2010) improved trad. of 2

ACKNOWLEDGMENTS

ISIR (CNRS – UPMC)

Nassim Aklil

Jean Bellot

Ken Caluwaerts

Dr. Laurent Dollé

Dr. Benoît Girard

Florian Lesaint

Pr. Olivier Sigaud

Guillaume Viejo

Univ. Sheffield

Pr. Kevin Gurney

Dr. Mark D. Humphries

Univ. Maryland / NIH-NIDA

Dr. Matthew R. Roesch

Pr. Geoffrey Schoenbaum

Financial support

FP6 IST 027189
European project

Learning under
Uncertainty Project

HABOT Project
Emergence(s) Program

REFERENCES (I)

- Baldassarre, G. (2002). A modular neural-network model of the basal ganglia's role in learning and selecting motor behaviors. Journal of Cognitive Systems Research, *3*(1), 5–13.
- Barto, A.G. (1995) Adaptive critics and the basal ganglia. In Houk, J.C., Davis, J.L. & Beiser, D.G. (Eds), Models of Information Processing in the Basal Ganglia. MIT Press, Cambridge, pp. 215–232.
- Benchenane, K., Peyrache, A., Khamassi, M., Wiener, S.I. and Battaglia, F.P. (2010). Coherent theta oscillations and reorganization of spike timing in the hippocampal-prefrontal network upon
- learning. Neuron, 66(6):921-36.
- Berns, G. S. and Sejnowski, T. J. (1996). How the basal ganglia make decision. In The neurobiology of decision making, A. Damasio, H. Damasio, and Y. Christen (eds), pages 101–113. Springer-Verlag, Berlin.
- Bertin, M., Schweighofer, N. and Doya, K. (2007). Multiple model-based reinforcement learning explains dopamine neuronal activity. Neural Networks, 20:668-675.
- Buffet, O. and Sigaud, O. (2008). Processus décisionnels de Markov en intelligence artificielle (volume 2). , Lavoisier, publisher.
- Caluwaerts, K., Staffa, M., N'Guyen, S. Grand, C., Dollé, L., Favre-Felix, A., Girard, B. and and Khamassi, M. (2012). A biologically inspired meta-control navigation system for the Psikharpax rat robot. Biomimetics & Bioinspiration, to appear..
- Daw, N. D. (2003). Reinforcement learning models of the dopamine system and their behavioral implications. Ph.D. thesis, Carnegie Mellon University, Pittsburgh, PA.
- Daw ND, Niv Y and Dayan P (2005). Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioral control. Nature Neuroscience, 8:1704-11.
- Devan BD and White NM (1999). Parallel information processing in the dorsal striatum: relation to hippocampal function. J Neurosci, 19(7):2789-98.

REFERENCES (II)

- Dollé L, Khamassi M, Girard B, Guillot A, Chavarriaga R (2008). Analyzing interactions between navigation strategies using a computational model of action selection. In Spatial Cognition VI, pp. 71-86, Springer LNCS 4095.
- Dollé, L. and Sheynikhovich,D. and Girard,B. and Chavarriaga,R. and Guillot,A. (2010). Path planning versus cue responding: a bioinspired model of switching between navigation strategies. Biological Cybernetics, 103(4):299-317.
- Doya K (2000). Complementary roles of basal ganglia and cerebellum in learning and motor control. Curr Opin Neurobiol, 10(6):732-9.
- Doya, K., Samejima, K., Katagiri, K., & Kawato, M. (2002) Multiple model-based reinforcement learning. Neural Computation, 14(6), 1347–1369.
- Doya,K.(2002).Metalearningand neuromodulation. NeuralNetw. 15, 495–506.
- Euston, D.R., Tatsuno, M., and McNaughton, B.L. (2007). Fast-forward playback of recent memory sequences in prefrontal cortex during sleep. Science 318, 1147–1150.
- Foster, D.J., and Wilson, M.A. (2006). Reverse replay of behavioural sequences in hippocampal place cells during the awake state. Nature 440, 680–683.
- Gupta, A.S., van der Meer, M.A.A., Touretsky, D.S. and Redish, A.D. (2010). Hippocampal Replay Is Not a Simple Function of Experience. Neuron 65, 695–705.
- Houk, J. C., Adams, J. L. & Barto, A. G. (1995). A Model of How the Basal Ganglia Generate and Use Neural Signals That Predict Reinforcement. In Houk et al. (Eds), Models of Information Processing in the Basal Ganglia (pp. 215-232). The MIT Press, Cambridge, MA.
- Joel, D., Niv, Y., & Ruppin, E. (2002). Actor–critic models of the basal ganglia: New anatomical and computational perspectives. Neural Networks, 15:535–547.
- Johnson, A., and Redish, A.D. (2007). Neural ensembles in CA3 transiently encode paths forward of the animal at a decision point. J. Neurosci. 27, 12176–12189.

REFERENCES (III)

- Keramati, M., Dezfouli, A., and Piray, P., Speed/Accuracy Trade-off between the Habitual and the Goal-directed Processes, PLOS Comput Bio, 7:5, 1-25 (2011).
- 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 Behav., 13, 131–148.
- Khamassi, M., Martinet, L.-E. & Guillot, A. (2006) Combining self-organizing maps with mixture of experts: Application to an Actor–Critic model of reinforcement learning in the basal ganglia. In Nolfi, S., Baldassare, G., Calabretta, R., Hallam, J., Marocco, D., Meyer, J.-A., Miglino, O. & Parisi, D. (Eds), From Animals to Animats 9, Proceedings of the Ninth International Conference on Simulation of Adaptive Behavior. Springer - Lecture Notes in Artificial Intelligence 4095, Springer, Berlin/Heidelberg, pp. 394–405.
- Khamassi, M., Mulder, A.B., Tabuchi, E., Douchamps, V. and Wiener S.I. (2008). Anticipatory reward signals in ventral striatal neurons of behaving rats. European Journal of Neuroscience, 28(9):1849-66.
- Khamassi, M., Lallée, S., Enel, P., Procyk, E. and Dominey P.F. (2011). Robot cognitive control with a neurophysiologically inspired reinforcement learning model. Frontiers in Neurorobotics, 5:1, doi:10.3389/fnbot.2011.00001.
- Kouneiher,F.,Charron,S.,andKoech- lin, E.(2009).Motivation and cognitive control in the human prefrontal cortex. Nat Neurosci,12, 939–945.
- Martinet, L.-E.; Sheynikhovich, D.; Benchenane, K. and Arleo, A. Spatial Learning and Action Planning in a Prefrontal Cortical Network Model. PLoS Comput Biol, 7 (5): e1002045, 2011.
- Montague, P. R., Dayan, P., & Sejnowski, T. J. (1996). A framework for mesencephalic dopamine systems based on predictive Hebbian learning. Journal of Neuroscience, 16, 1936–1947.

REFERENCES (IV)

- Morris, G., Nevet, A., Arkadir, D., Vaadia, E., Bergman, H. (2006). Midbrain dopamine neurons encode decisions for future action. Nat Neurosci 9(8):1057–1063.
- Packard MG and Knowlton BJ (2002). Learning and memory functions of the Basal Ganglia. Annu Rev Neurosci, 25:563-93.
- Pearce JM, Roberts AD and Good M (1998). Hippocampal lesions disrupt navigation based on cognitive maps but not heading vectors. Nature, 396(6706):75-7.
- Peyrache, A., Khamassi, M., Benchenane, K., Wiener, S.I. and Battaglia, F.P. (2009). Replay of rule-learning related neural patterns in the prefrontal cortex during sleep. Nature Neuroscience, 12(7):919-26.
- Quilodran,R.,Rothe,M.,and Procyk,E. (2008). Behavioral shifts and action valuation in the anterior cingulate cortex. Neuron 57, 314–325.
- Rescorla RA, Wagner AR. A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. In: Classical Conditioning II: Current Research and Theory (Eds Black AH, Prokasy WF) New York: Appleton Century Crofts, pp. 64-99, 1972.
- Roesch, M.R., Calu, D.J., Schoenbaum, G. (2007). Dopamine neurons encode the better option in rats deciding between differently delayed or sized rewards. Nat Neurosci 10(12):1615–1624.
- Schweighofer N, Doya K (2003) Meta-learning in reinforcement learning. Neural Netw 16:5-9.
- Schultz, W., Apicella, P. & Ljungberg, T. (1993). Responses of Monkey Dopamine Neurons to Reward and Conditioned Stimuli During Successive Steps of Learning a Lelayed Response Task. Journal of Neuroscience, 13(3):900-913.
- Schultz, W., Dayan, P., & Montague, P. R. (1997). A neural substrate of prediction and reward. Science, 275, 1593–1599.
- Sigaud, O. and Buffet, O. (2010). Markov Decision Processes in Artificial Intelligence. iSTE - Wiley, publisher.

REFERENCES (V)

- Suri RE and Schultz W (1999). A neural network model with dopamine-like reinforcement signal that learns a spatial delayed response task. Neuroscience, 91(3):87190.
- Suri, R. E., & Schultz, W. (2001). Temporal difference model reproduces anticipatory neural activity. Neural Computation, 13, 841–862.
- Sutton, R. S., & Barto, A. G. (1998). Reinforcement learning: An introduction. Cambridge, MA: The MIT Press.
- Sutton RS (1990). Integrated architectures for learning, planning, and reacting based on approximating dynamic programming. In Seventh International Machine Learning Workshop, pages 21624. Morgan Kaufmann, San Mateo, CA.
- Wilson, M.A., and McNaughton, B.L. (1994). Reactivation of hippocampal ensemble memories during sleep. Science 265, 676–679.

LECTURES & COMMENTARIES

- Balleine, B. (2005). Prediction and control: Pavlovian-instrumental interactions and their neural bases. Lecture at OCNC 2005: http://www.irp.oist.jp/ocnc/2005/lectures.html#Balleine.
- Daw, N.D. (2007). Dopamine: at the intersection of reward and action. News and views in Nature Neuroscience, 9(8).
- Niv, Y., Daw, N.D. and Dayan, P. (2006). Choice values. New and views in Nature Neuroscience, 9(8).