Artificial Neural Networks and RL

The role of exploration, novelty, and surprise in RL

Wulfram Gerstner

EPFL, Lausanne, Switzerland

Objectives for today:

- understand surprise
- understand difference of novelty and surprise
- use of surprise to modulate learning rate
- use of novelty to guide exploration

Novelty and Surprise

Q1: What is novelty?

Q2: What is surprise?

Q3: What is the difference between the two?

Q4: Why are they useful?

Q5: Why should we talk about it in an RL class?

Enjoy the images!

Novelty is not Surprise

Surprise is against models (beliefs)

Novelty and Surprise

Q3: *What is the difference between the two?*

First answer – **novelty and surprise are not the same.**

Second answer (more precise):

Surprise is ‘against beliefs’ or ‘against expectations’ whereas novelty is not.

Novelty and Surprise

Surprise is ‘against expectations’: an example

…

Artificial Neural Networks and RL

Wulfram Gerstner
EPFL, Lausanne, Switzerland

The role of exploration, novelty, and surprise in RL

1. Definitions of Novelty and Surprise (tabular environment)

Novelty in a tabular environment: discrete states

events = states s (e.g., one image). Total number is |s|

Novelty *n*:

1) count events of type s up to time t: $C^t(s)$

2) a higher count gives lower novelty.

3) the agent has spent a time t in the environment

4) the empirical observation frequency is $p_N(s) = \frac{C^t(s)+1}{t+|s|}$

Definition: The **'Novelty' of a state s** at time t is

$$n_t(s) = -\log p_N(s)$$

Surprise in a tabular environment: discrete states and actions

events = transitions (s,a→s') given action a in state s.

Surprise S**:**

1) count events of type (s,a→s') up to time t: $C^t(s, a \to s')$
2) a higher count gives lower surprise.
3) the agent has spent a time t in the environment
4) the empirical observation frequency is

$$p^t\ (s_{t+1} = s'|s_t, a_t) = \frac{C^t(s, a \to s') + 1}{\tilde{C}^t(s, a) + |s|}$$

Definition: The **'Surprise' of a transition is**

$$S_{BF}^{t+1}(s') = \frac{prior}{p_s^t(s_{t+1} = s'|s_t, a_t)}$$

Bayes Factor Surprise

Definitions of Novelty and Surprise

Q1: What is novelty?

Definition: The **'Novelty' of a state s** is

$$n^t(s) = -\log p_N(s)$$

Q2: What is surprise?

Definition: The **'Surprise' of a transition is**

$$S_{BF}^{t+1}(s') = \frac{prior}{p_s^t(s_{t+1} = s' \mid s_t, a_t)}$$

There are 17 different definitions of surprise.
This here is the Bayes-Factor surprise.

Modirshanechi et al. (2022)

Artificial Neural Networks and RL

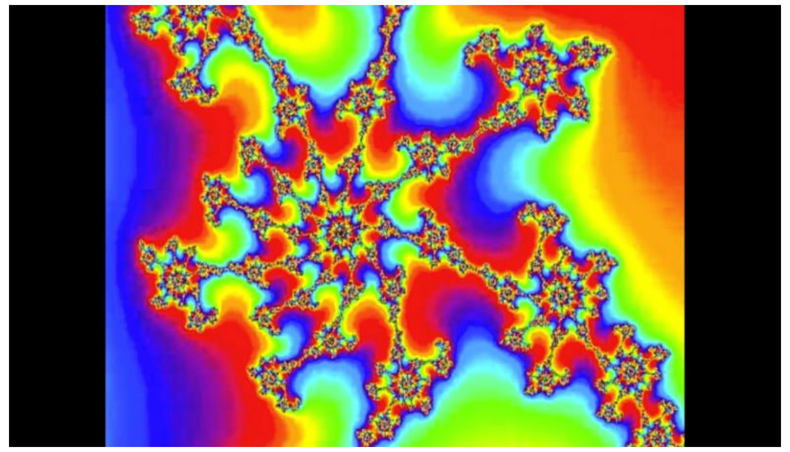
Wulfram Gerstner
EPFL, Lausanne, Switzerland

The role of exploration, novelty, and surprise in RL

1. **Definitions of Novelty and Surprise (tabular environment)**
2. **Why is Surprise useful?**

When are we surprises?

3 9 7 3 9 7 3 9 7 3 9 7 3 9 4 3 9 7



Surprise against expectations from your current belief

- Expectations arise from models of the world
- We always make models
- We know that the models are not perfect
- **Surprise enables us to adapt the models**

→ **Hypothesis:**

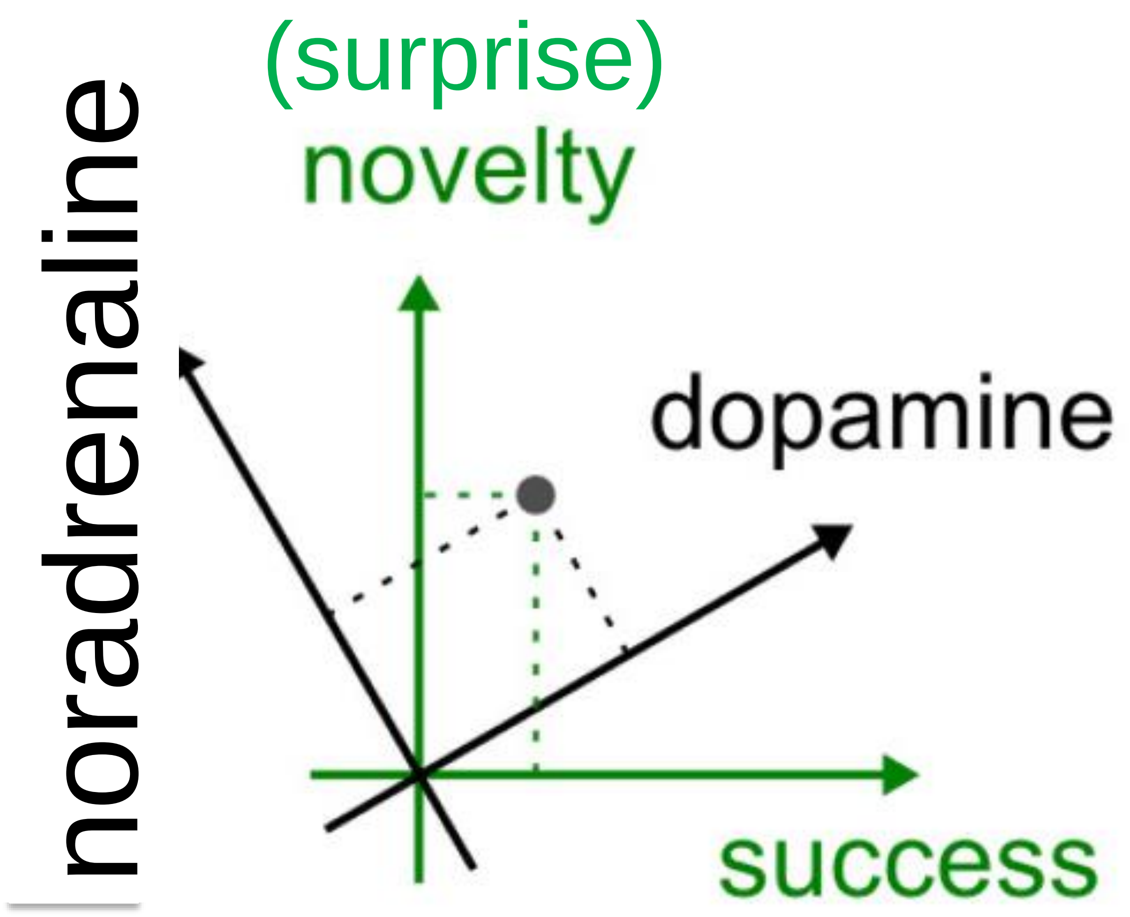
Surprise boosts plasticity (3rd factor)/ increases the learning rate

Note: no reward!!!!

Review: Neuromodulators

- 4 or 5 neuromodulators
- near-global action
- internally created signals

Dopamine/reward/TD:
Schultz et al., 1997,
Schultz, 2002

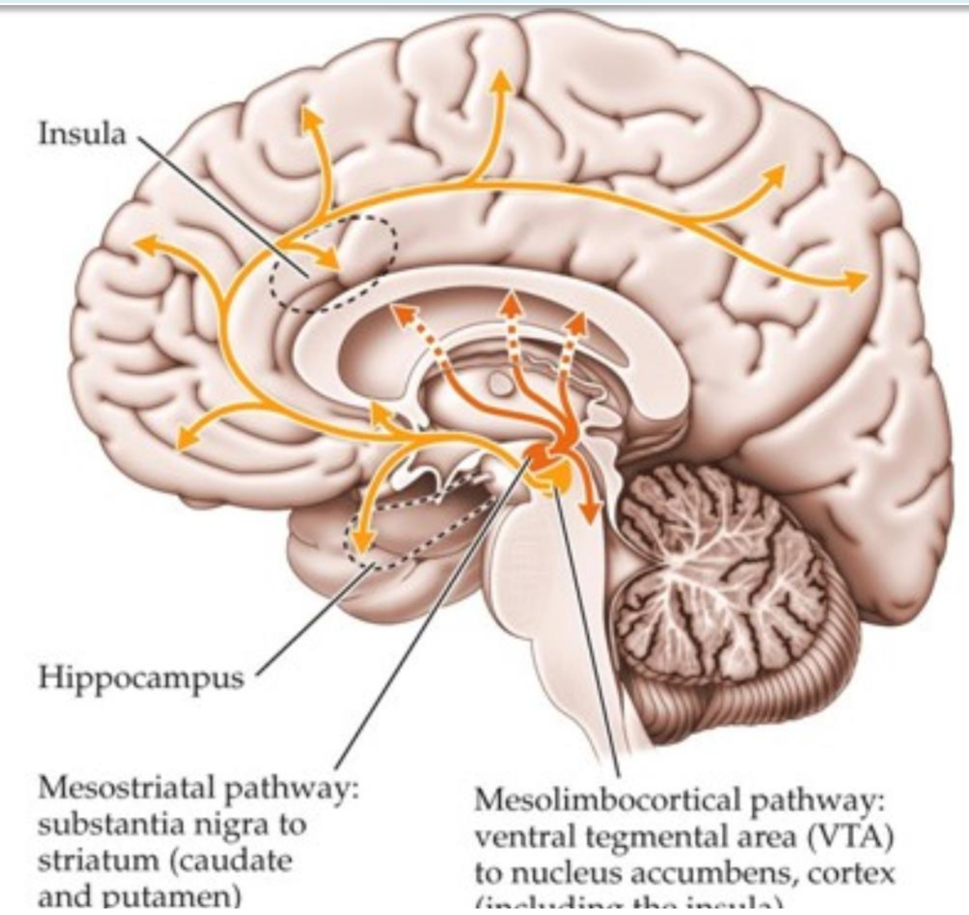


$[r+\gamma V(s')-V(s)]$

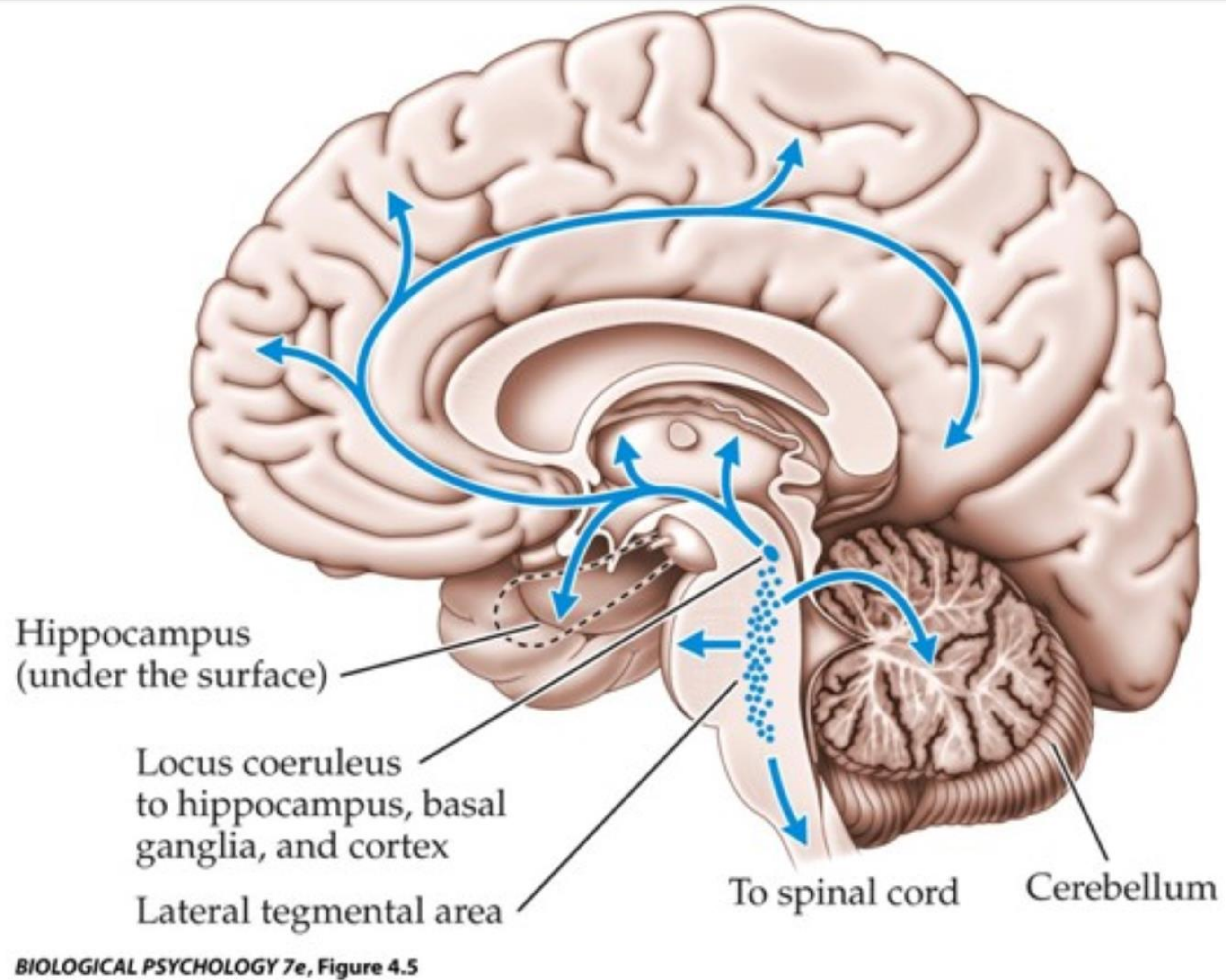
Image:
Fremaux and Gerstner, Frontiers (2016)

Image: Biological Psychology, Sinauer

Dopamine (DA)



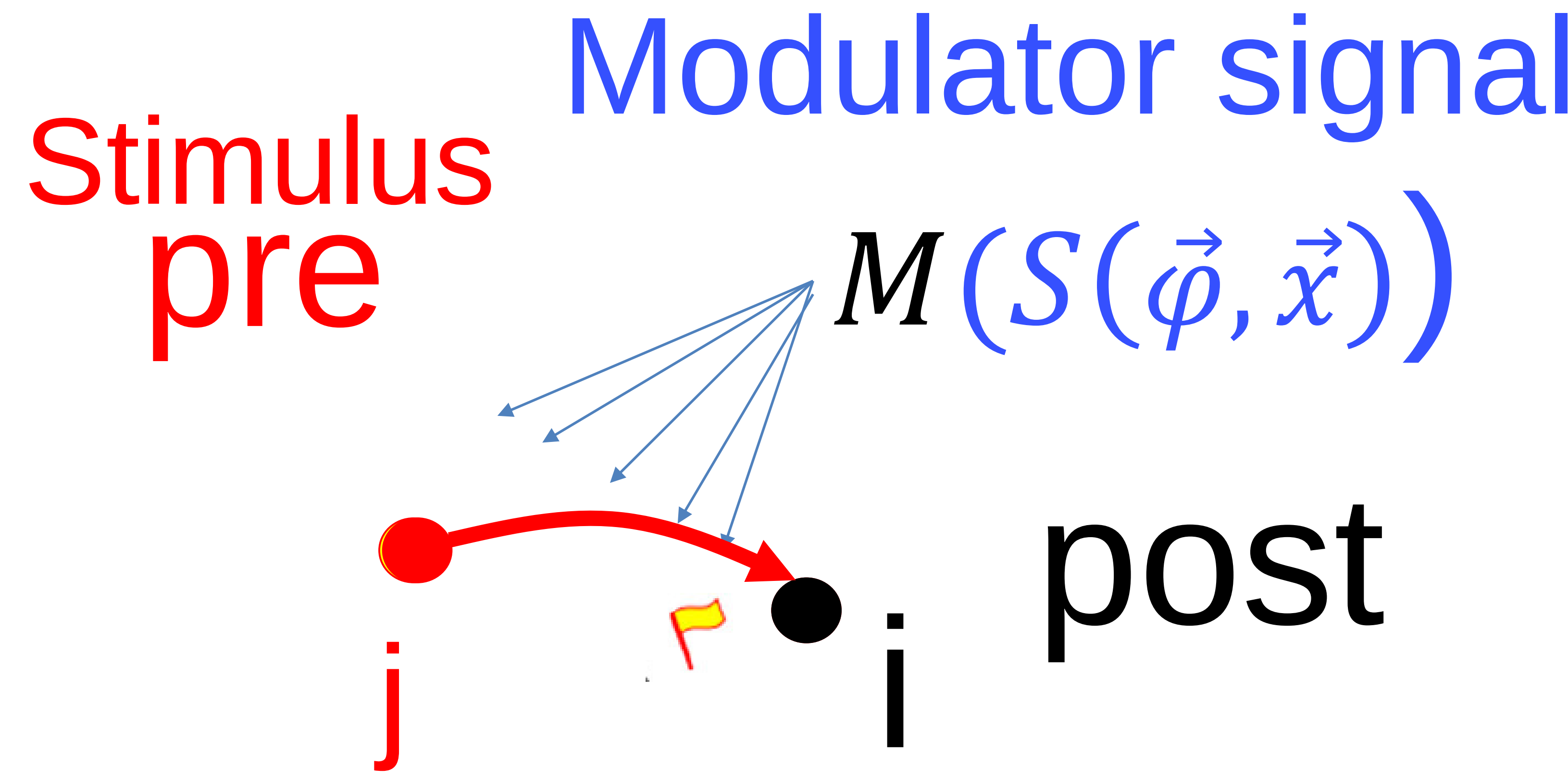
Noradrenaline (NE)



Review: Formalism of Three-factor rules with eligibility trace

x_j = activity of presynaptic neuron

φ_i = activity of postsynaptic neuron



Step 1: co-activation sets eligibility trace

$$\Delta z_{ij} = \eta \; f(\varphi_i)\, g(x_j)$$

Step 2: eligibility trace decays over time

$$z_{ij} \leftarrow \lambda\, z_{ij}$$

Step 3: eligibility trace translated into weight change

$$\Delta w_{ij} = \eta\, M\big(S(\vec{\varphi},\vec{x})\big) z_{ij}$$

$M(S)$:
- TD-error
- surprise

Artificial Neural Networks and RL

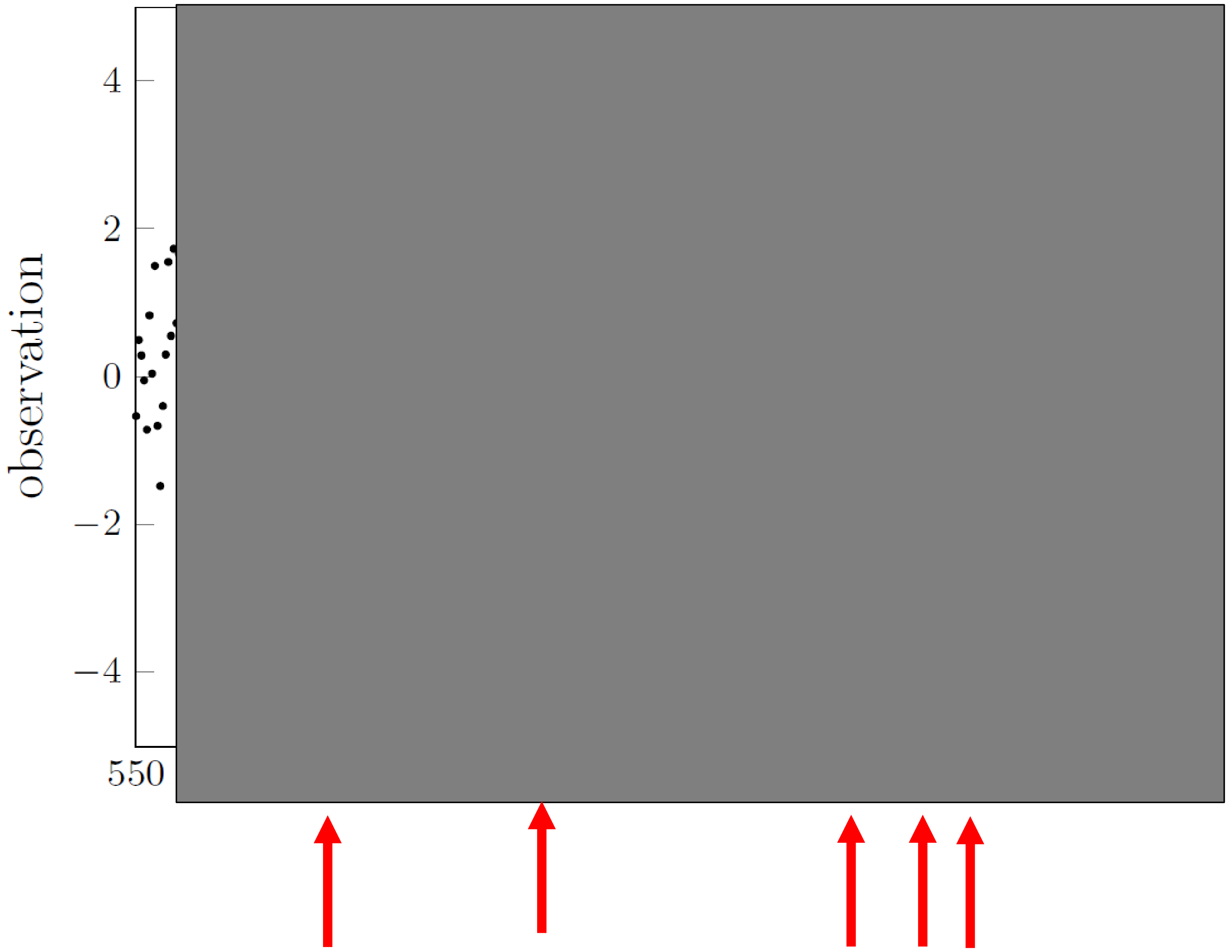
Wulfram Gerstner

EPFL, Lausanne, Switzerland

The role of exploration, novelty, and surprise in RL

1. Definitions of Novelty and Surprise (tabular environment)
2. Why is Surprise useful?
3. Change-point detection by Bayes-Surprise

Surprise boosts plasticity in volatile environments



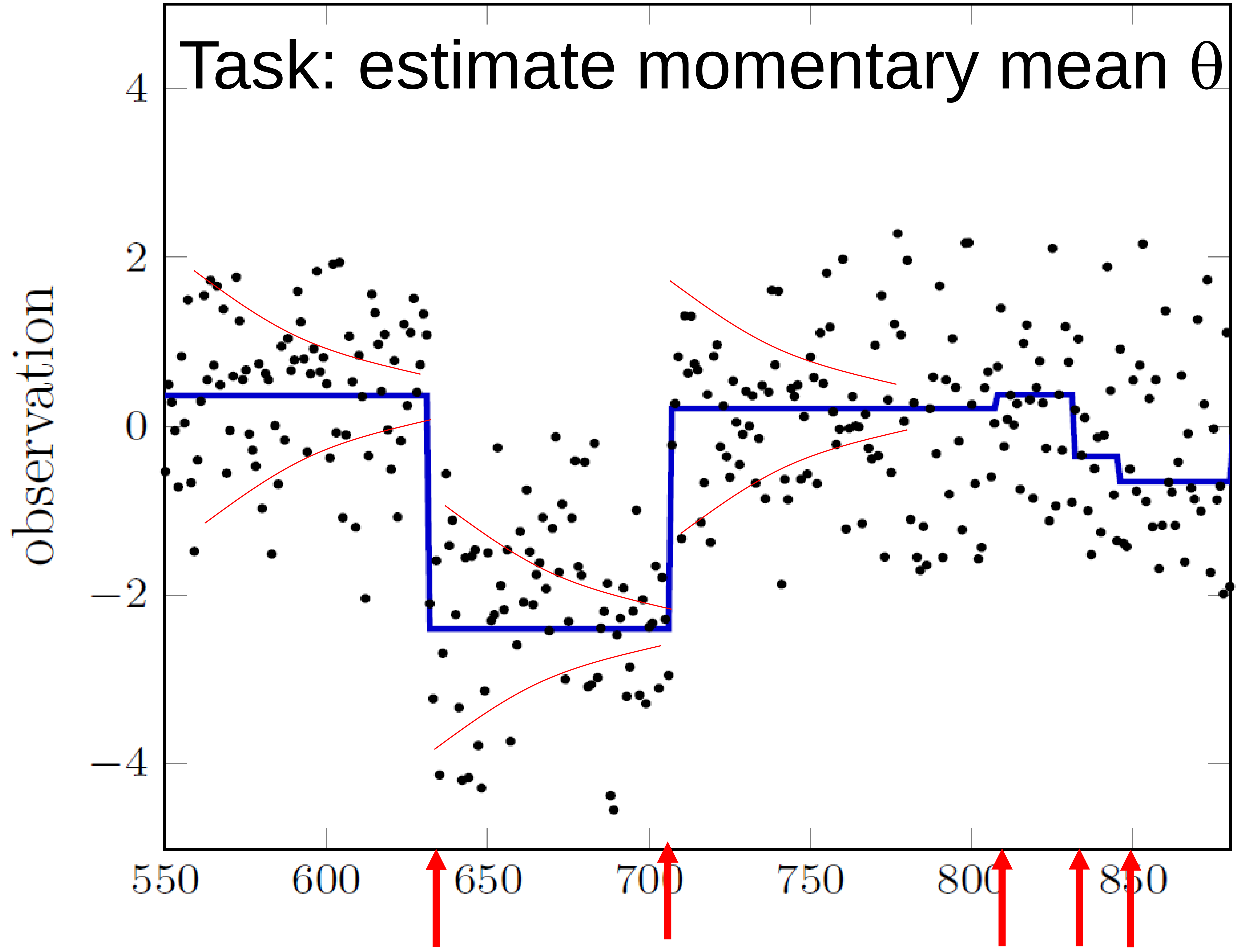
Volatile environment:
abrupt changes with small probability
→ 'change points'

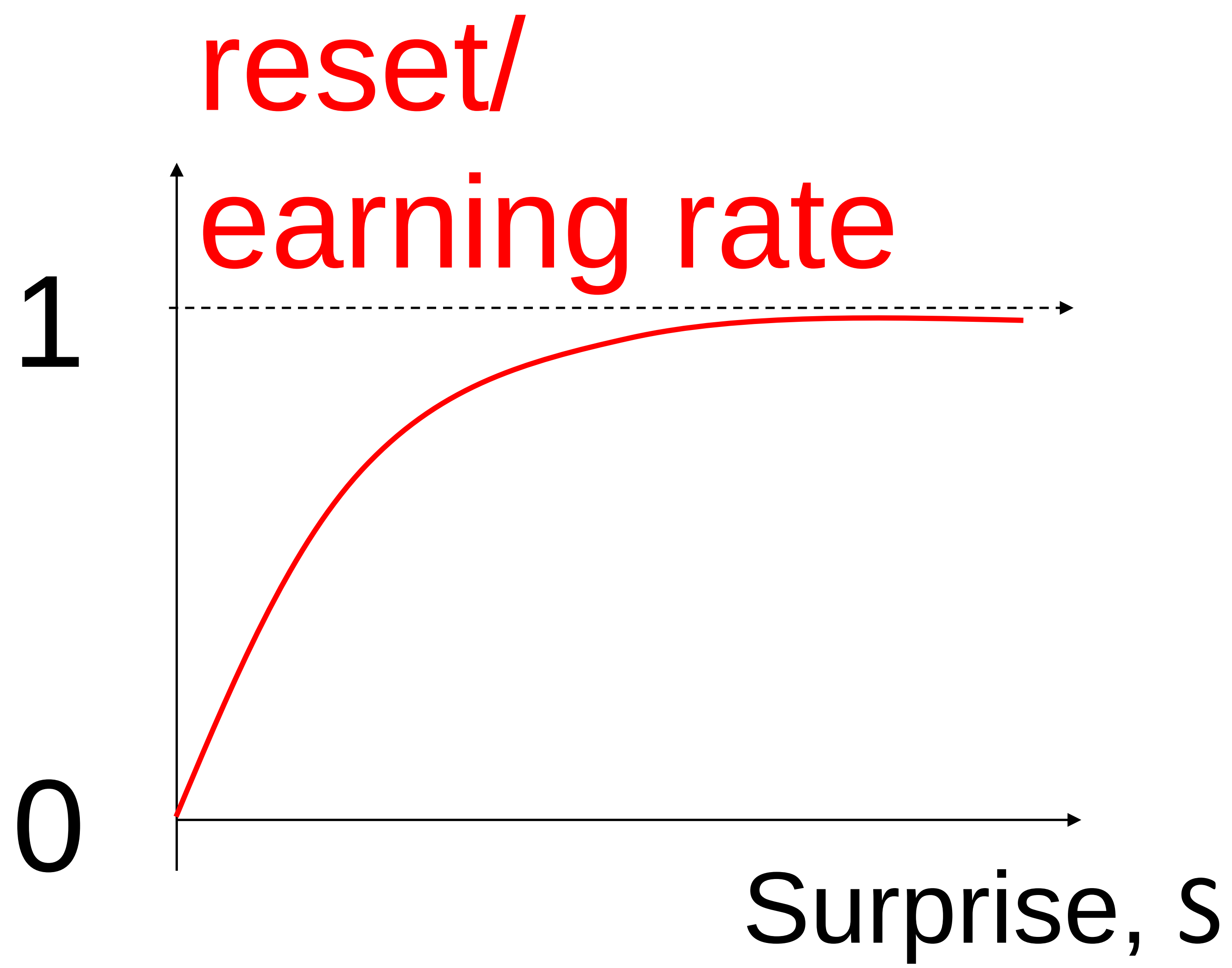
→ you have to reset model after a change point

generative model = nonstationary stochastic process

here: - mean of Gaussian is fixed for many steps
- mean jumps at 'change points' : probability << 1
- variance is fixed
- task is to estimate **momentary mean** of Gaussian

Surprise boosts plasticity in volatile environments





in volatile environment, best approach (Bayesian):

- reset your belief to prior, if observation does not make sense
- plasticity of system must increase if ‘surprising observation’

Surprise boosts plasticity in volatile environments

$$S_{\mathrm{BF}}(y_{t+1};\pi^{(t)}) = \frac{P(y_{t+1};\pi^{(0)})}{P(y_{t+1};\pi^{(t)})}.$$

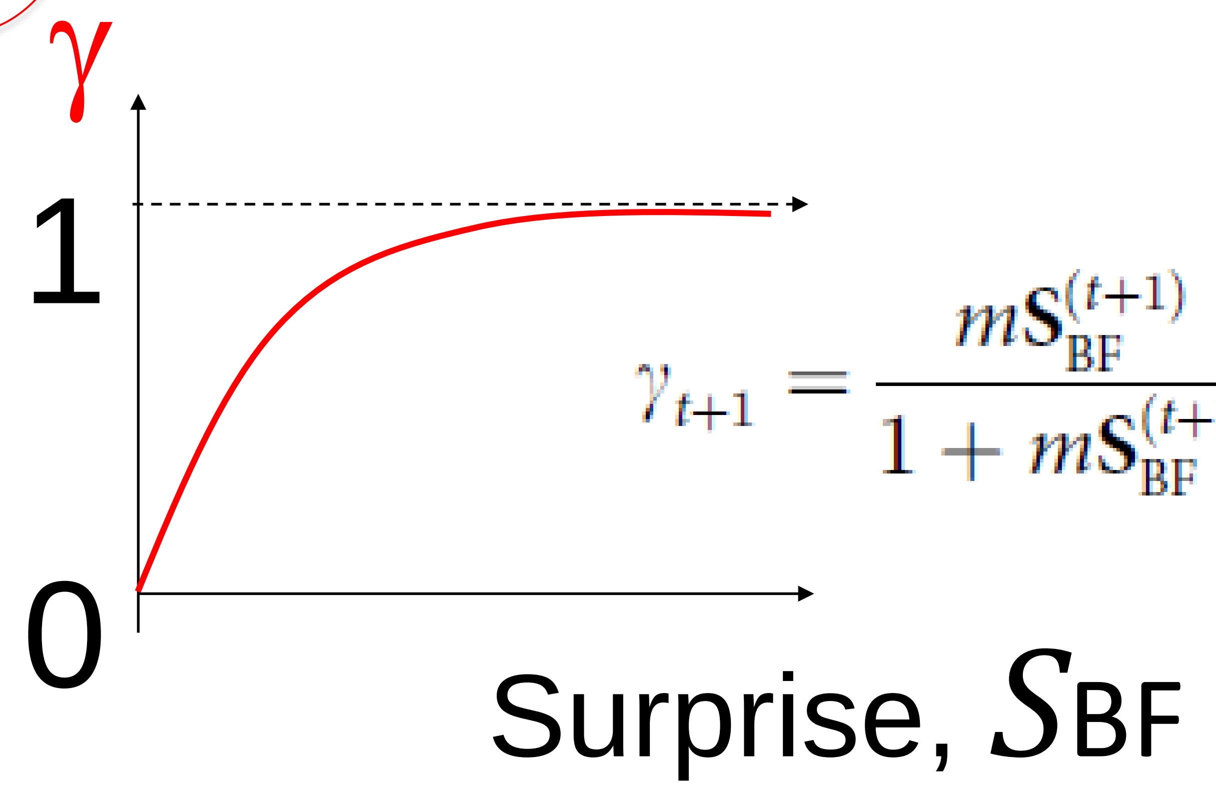
Probability of observation y under prior belief $\pi^{(0)}$

Probability of observation y under current belief $\pi^{(t)}$

→reset your belief to prior, if observation y does not make sense

$$\pi^{\text{new}}(\theta) = (1-\gamma)\,\pi^{\text{integration}}(\theta|y^{\text{new}},\pi^{\text{old}}) + \gamma\,\pi^{\text{reset}}(\theta|y^{\text{new}},\pi^{(0)}).$$

→'exact Bayesian inference'
in volatile environment modulates
update with factor γ



$$\gamma_{t+1} = \frac{m S_{\mathrm{BF}}^{(t+1)}}{1 + m S_{\mathrm{BF}}^{(t+1)}} \in [0,1],$$

Surprise boosts plasticity in volatile environments

$$S_{\mathrm{BF}}(y_{t+1}; \pi^{(t)}) = \frac{P(y_{t+1}; \pi^{(0)})}{P(y_{t+1}; \pi^{(t)})}.$$

Probability of observation y under prior belief $\pi^{(0)}$

Probability of observation y under current belief $\pi^{(t)}$

→reset your belief to prior, if observation y does not make sense

Exact update rule not implementable, but

Bayes-Factor Surprise plays crucial role in approximate methods:

- Particle Filter with N particles,
- Message-Passing with N messages,
- Published approximations

V. Liakoni et al., ***Neural Computation 2021***

Artificial Neural Networks and RL

Wulfram Gerstner

EPFL, Lausanne, Switzerland

The role of exploration, novelty, and surprise in RL

1. Definitions of Novelty and Surprise (tabular environment)
2. Why is Surprise useful?
3. Change-point detection by Bayes-Factor Surprise
4. Why is Novelty useful?

Review: TD-learning in the general sense

$$Q(s,a) = \sum_{s'} P^a_{s\to s'}\left[R^a_{s\to s'} + \gamma \sum_{a'} \pi(s',a') Q(s',a') \right]$$

SARSA

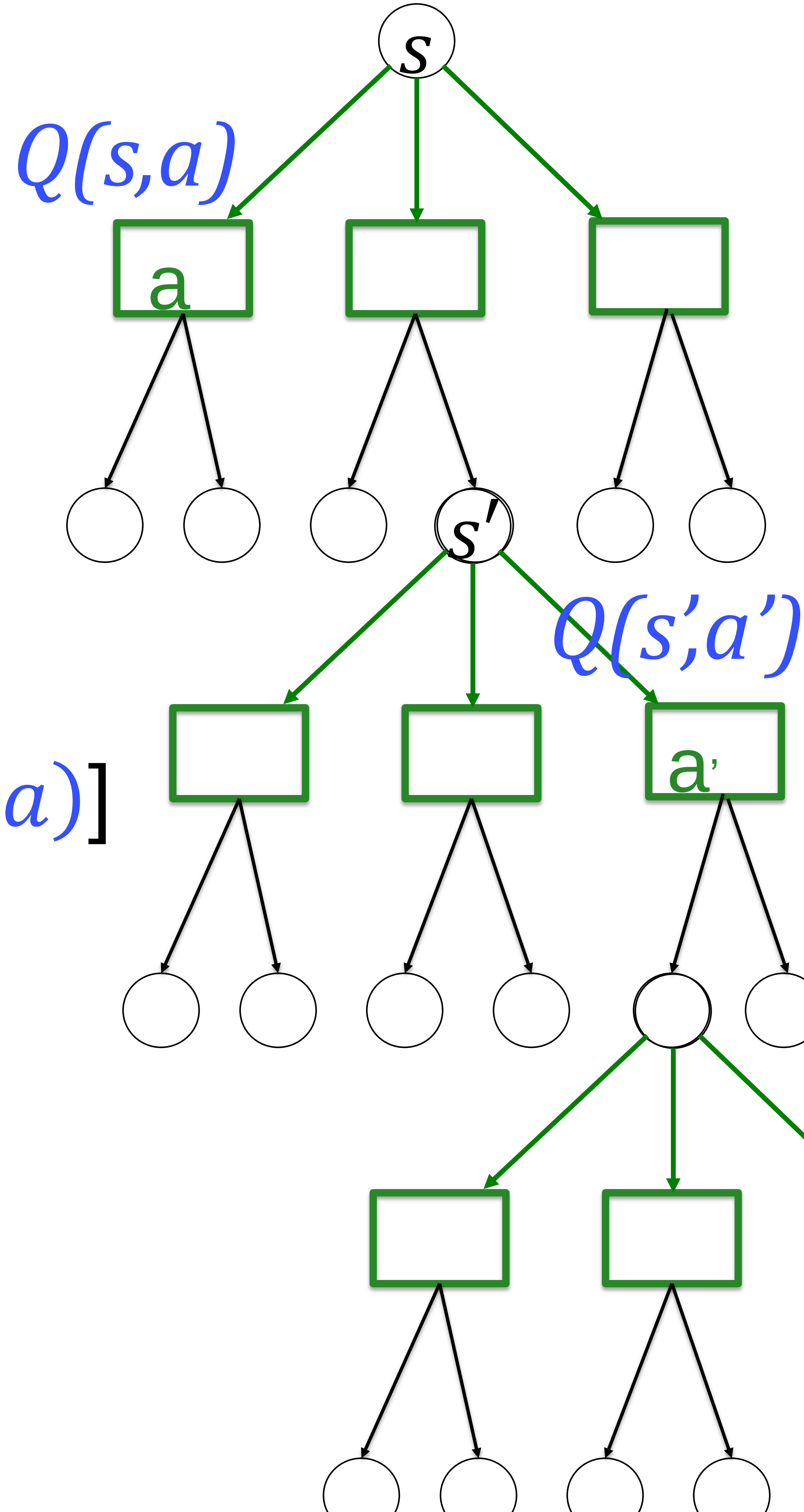
$$\Delta Q(s,a) = \eta[r_t + \gamma Q(s',a') - Q(s,a)]$$

Expected SARSA

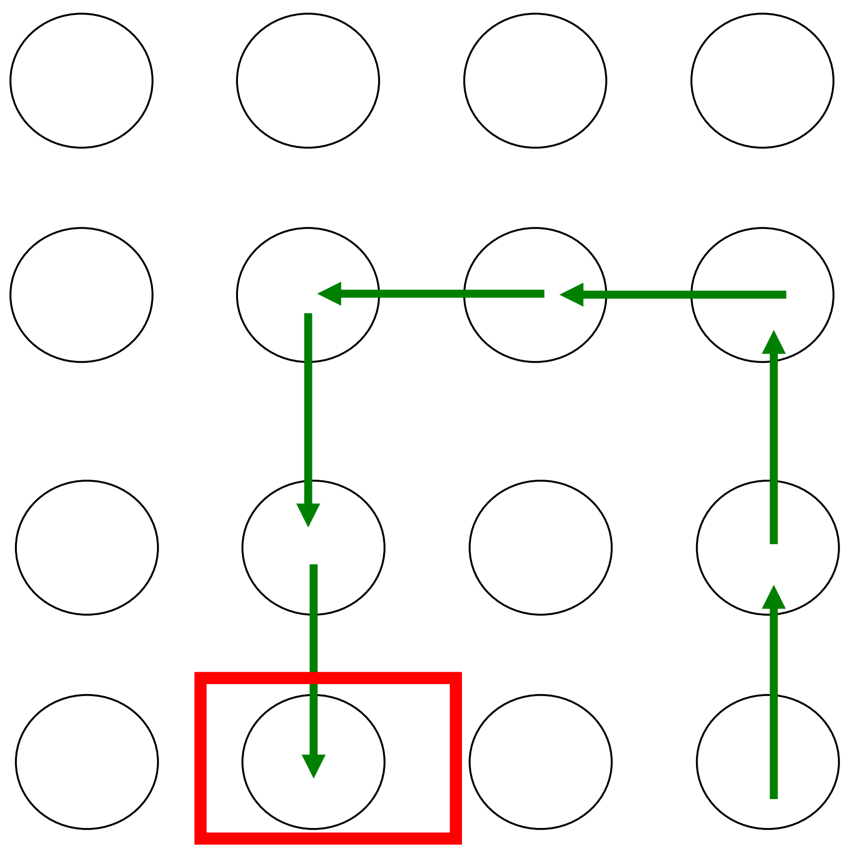
$$\Delta Q(s,a) = \eta[r_t + \gamma\{\textstyle\sum_{a'} \pi(s',a') Q(s',a')\} - Q(s,a)]$$

Q-learning

$$\Delta Q(s,a) = \eta[r_t + \gamma \max_{a'} Q(s',a') - Q(s,a)]$$



Review: Eligibility Traces, SARSA(λ)

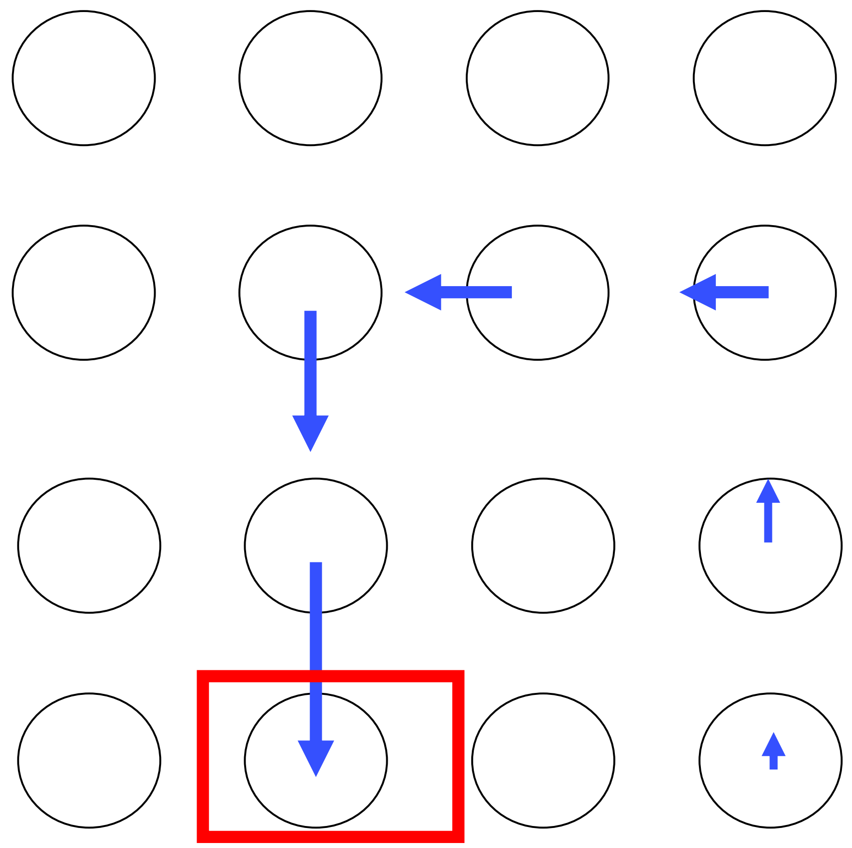


Idea:

- keep memory of previous state-action pairs
- memory decays over time
- update eligibility trace for **all** state-action pairs

$e(s,a) \leftarrow \lambda\, e(s,a)$ decay of **all** traces

$e(s,a) \leftarrow e(s,a) + 1$ if action a chosen in state s



- update **all** Q-values at **all time steps** t:

$$\Delta Q(s,a) = \eta\ \underbrace{[r_t + \gamma Q(s_{t+1},a_{t+1}) - Q(s_t,a_t)]}_{\text{RPE = TD error } \delta_t}\ e(s,a)$$

Note: λ=0 gives standard SARSA

Review: Model-based versus Model-free

Model-based

- learns model of environment 'transition matrix'
- knows 'rules' of game
- planning ahead is possible
- can update Bellman equation in 'background' without action
- can simulate action sequences (without taking actions)
- is not

Model-free

- does not
- does not
- cannot plan ahead
- cannot
- cannot
- Eligibility traces and V-values keep memory of past
- completely online, causal, forward in time.

Reward-based learning versus Novelty-based learning

rewards r_t

Q-values $Q_R^{(t)}(s,a)$

Bellman eq.
estimation/update

Model-based → prioritized sweeping → $Q_{MB,R}^{(t)}(s,a)$

Model-free → eligibility traces → $Q_{MF,R}^{(t)}(s,a)$

novelty n_t

Q-values $Q_N^{(t)}(s,a)$

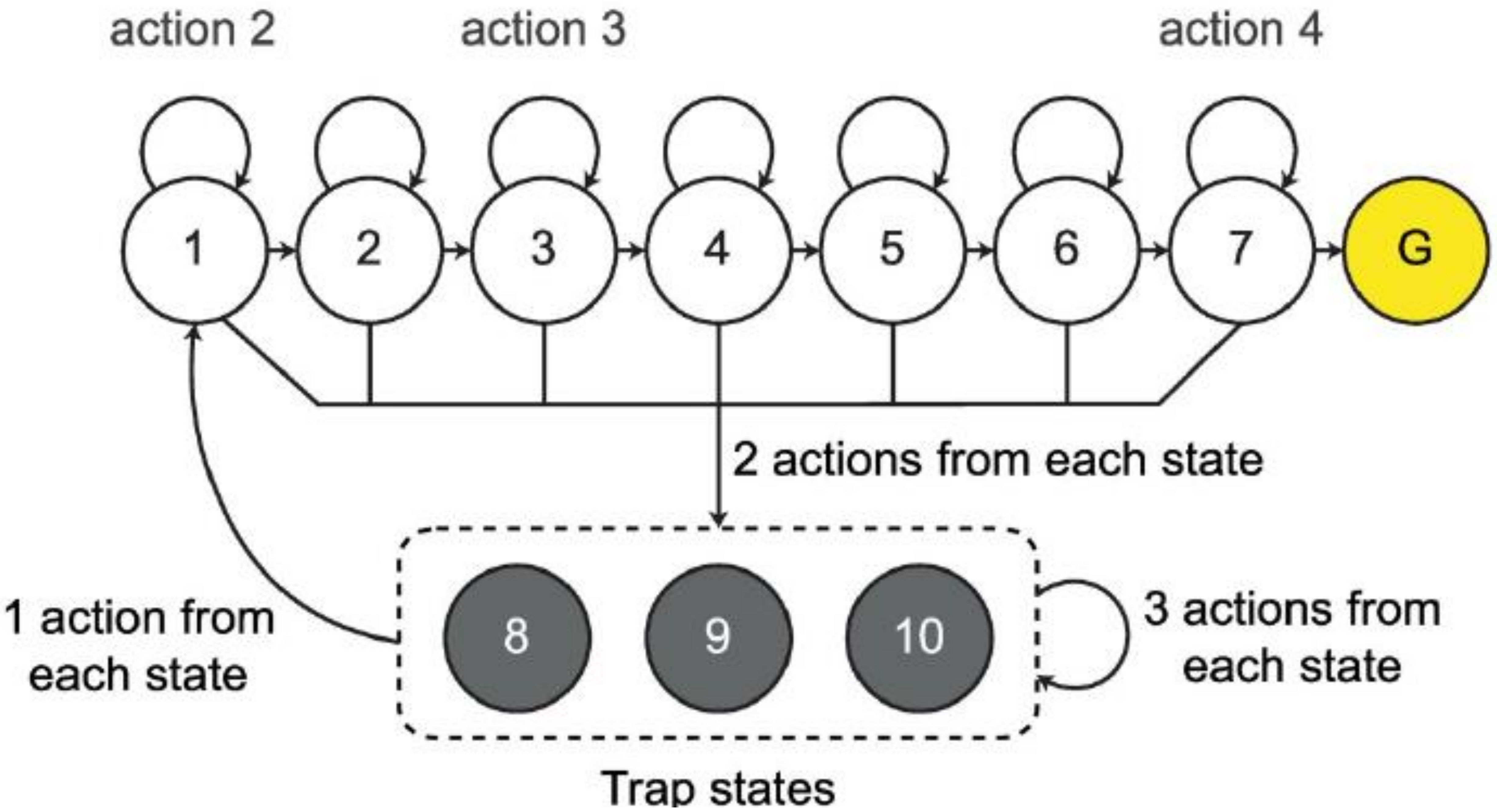
Bellman eq.
estimation/update

Model-based → prioritized sweeping → $Q_{MB,N}^{(t)}(s,a)$

Model-free → eligibility traces → $Q_{MF,N}^{(t)}(s,a)$

Initial exploration of an environment

Environment with 10 states (+ goal)
4 actions per state



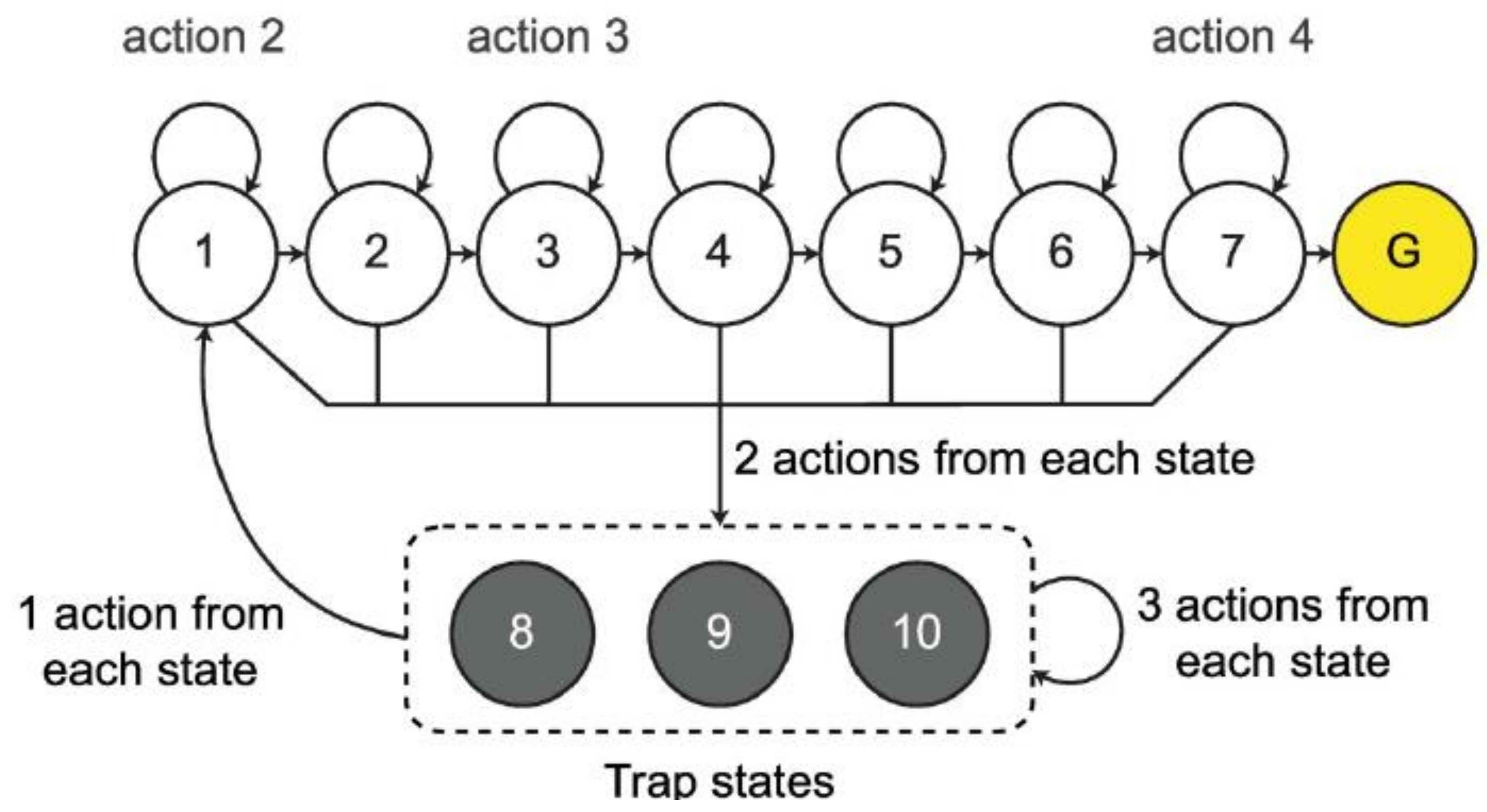
Actions are deterministic.
Fixed random assignment.

Start in state 1:
With random policy, how many actions on average before finding goal?
[ ] 100-500
[ ] 1000 – 5000
[ ] more than 10000

Improve exploration of an environment

Focus on 1st episode, before any reward.

Improve exploration! Solutions?



1. Optimistic initialization?

Initialize $Q_R(s,a) = 10$ for all s,a

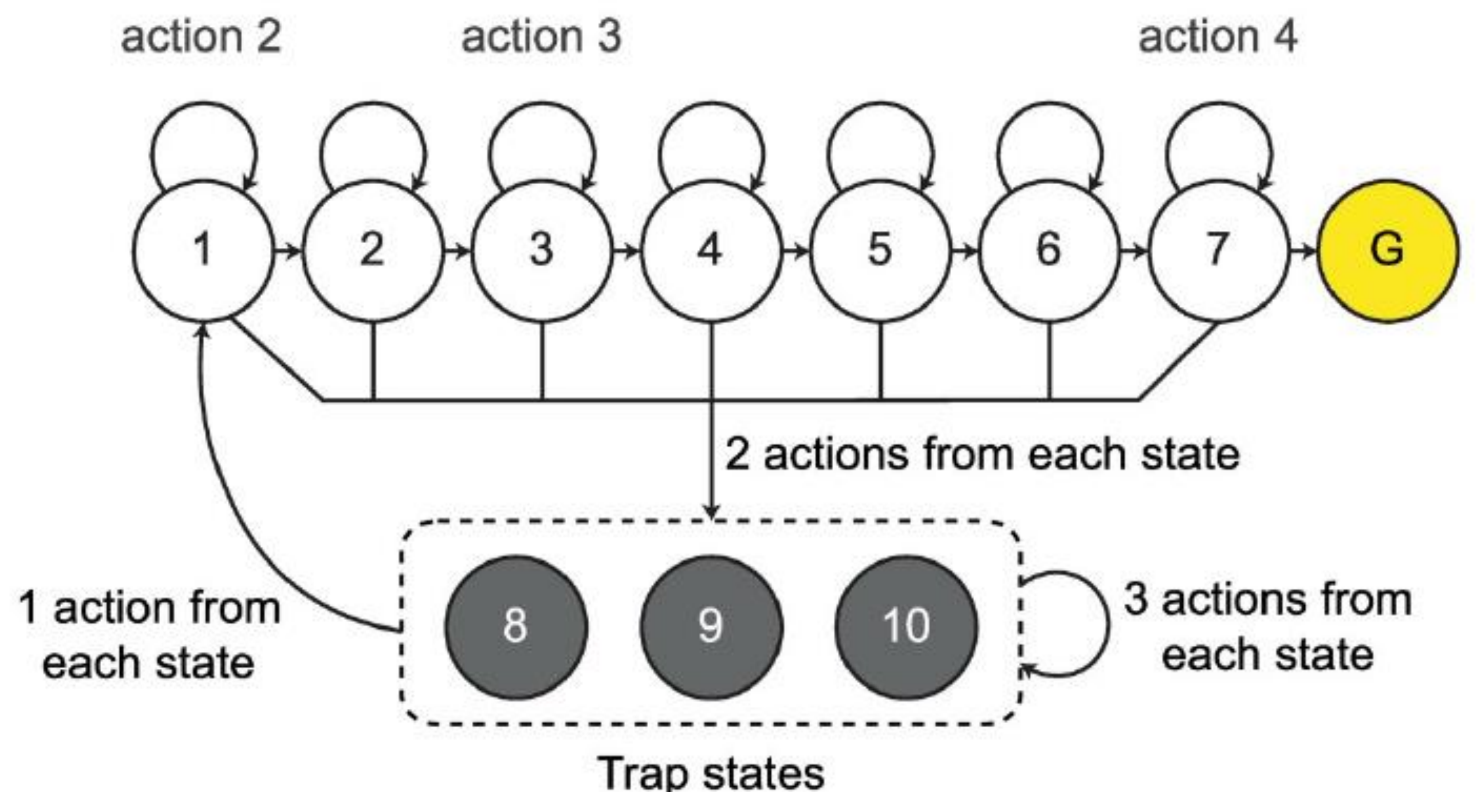
$$\Delta Q_R(s,a) = \eta[r_t + \gamma \max_{a'} Q_R(s',a') - Q_R(s,a)]$$

→ Possible but comparatively slow.
→ Does not generalize well for episode 2.

Novelty encourages exploration of an environment

Focus on 1st episode, before any reward.

Improve exploration! Solutions?



2. Novelty at time t is n_t

Novelty Prediction Error (NPE)

$$\Delta Q_N(s,a) = \eta[n_t + \gamma \max_{a'} Q_N(s',a') - Q_N(s,a)]$$

→ Separate Q-value for novelty!

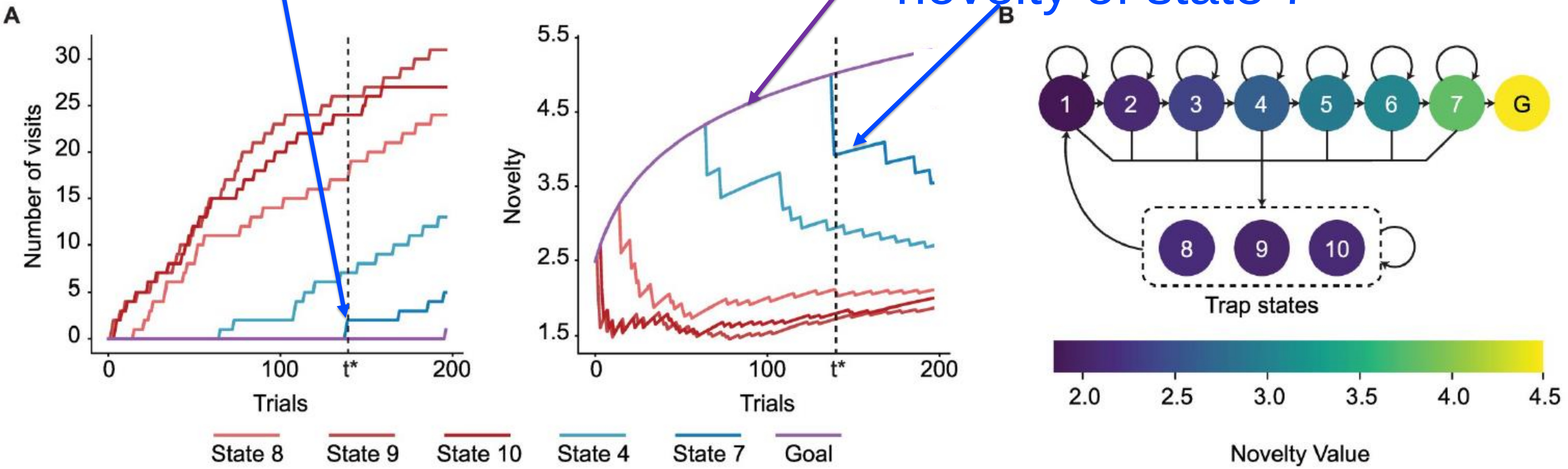
Novelty encourages exploration of an environment

Focus on 1st episode, before any reward; with some policy

first encounter of state 7

novelty of goal

novelty of state 7



→ use novelty values $Q_N^{(t)}(s,a)$ for action policy!

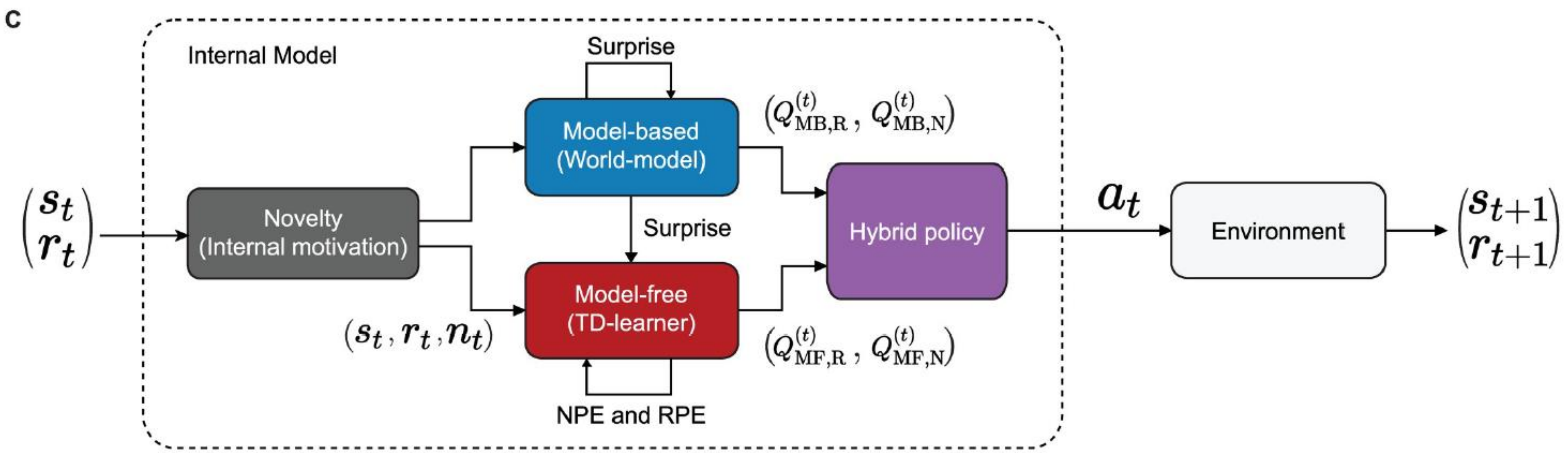
Artificial Neural Networks and RL

Wulfram Gerstner
EPFL, Lausanne, Switzerland

The role of exploration, novelty, and surprise in RL

1. Definitions of Novelty and Surprise (tabular environment)
2. Why is Surprise useful?
3. Change-point detection by Bayes-Factor Surprise
4. Why is Novelty useful?
5. Hybrid Model with Novelty, Surprise, and Reward

Hybrid model with separate paths for Novelty and Reward (learning rate controlled by Surprise)



RPE = $[r_t + \gamma \max_{a'} Q_R(s', a') - Q_R(s, a)]$

NPE = $[n_t + \gamma \max_{a'} Q_N(s', a') - Q_N(s, a)]$

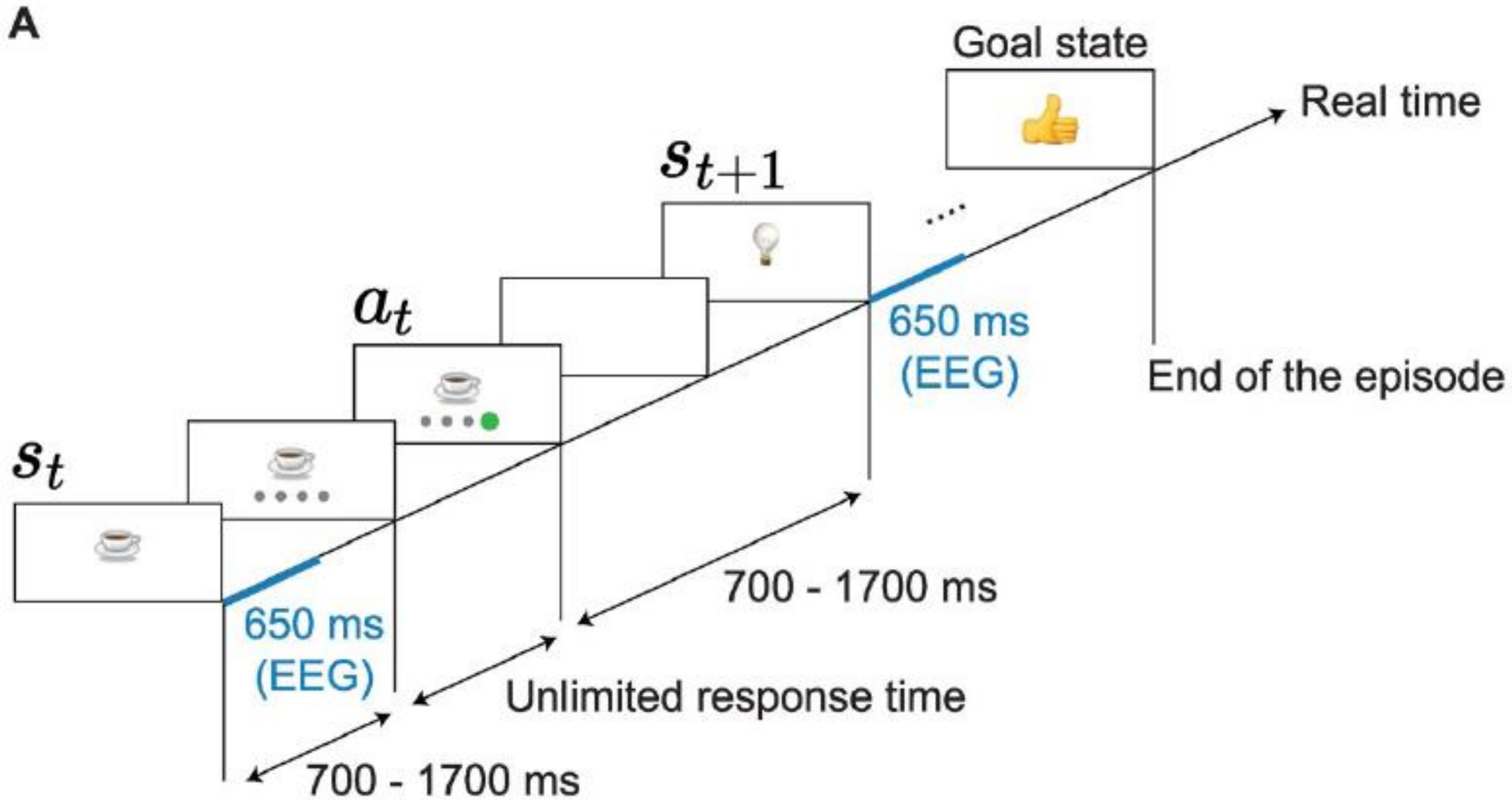
Artificial Neural Networks and RL

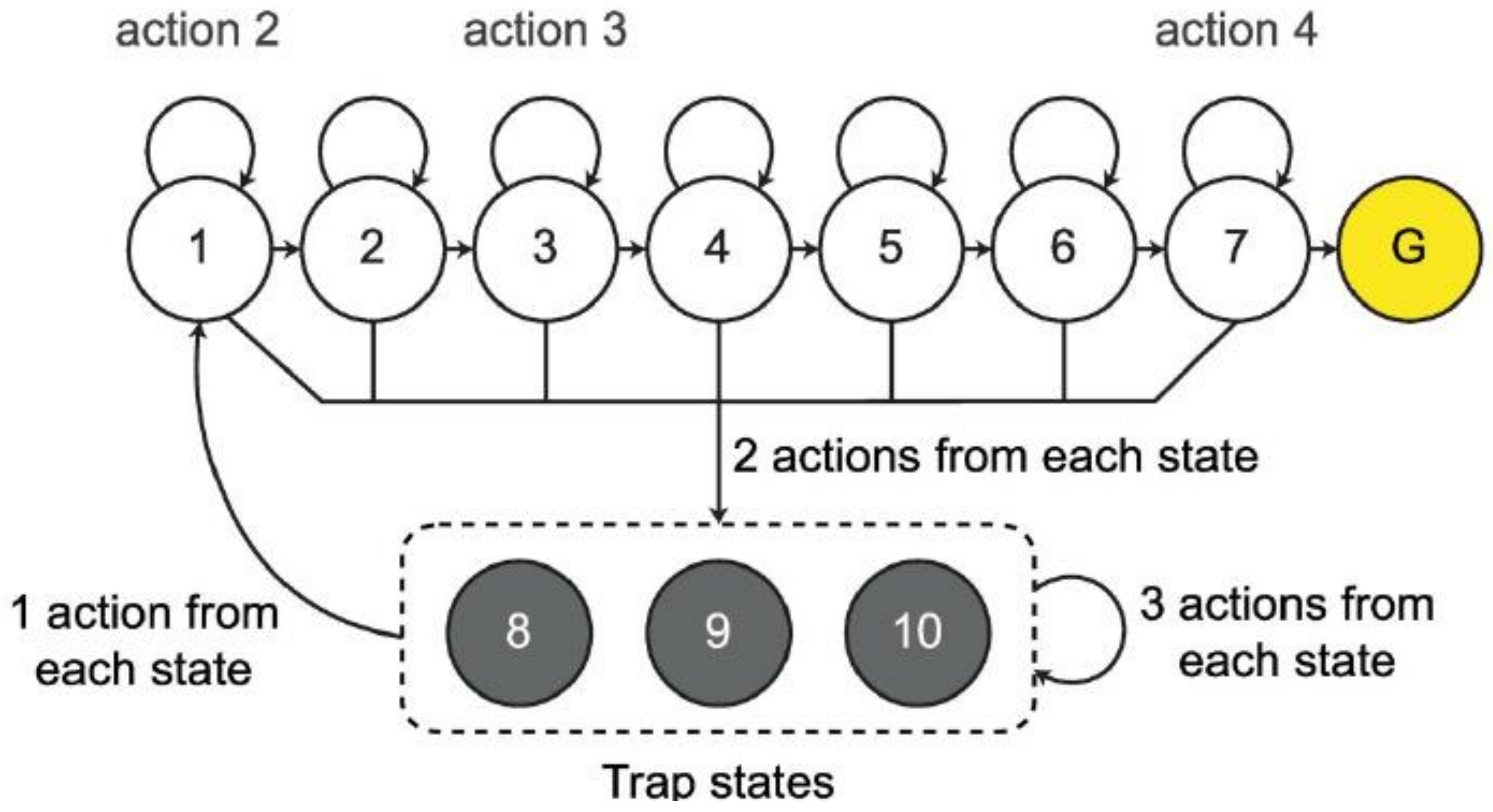
Wulfram Gerstner
EPFL, Lausanne, Switzerland

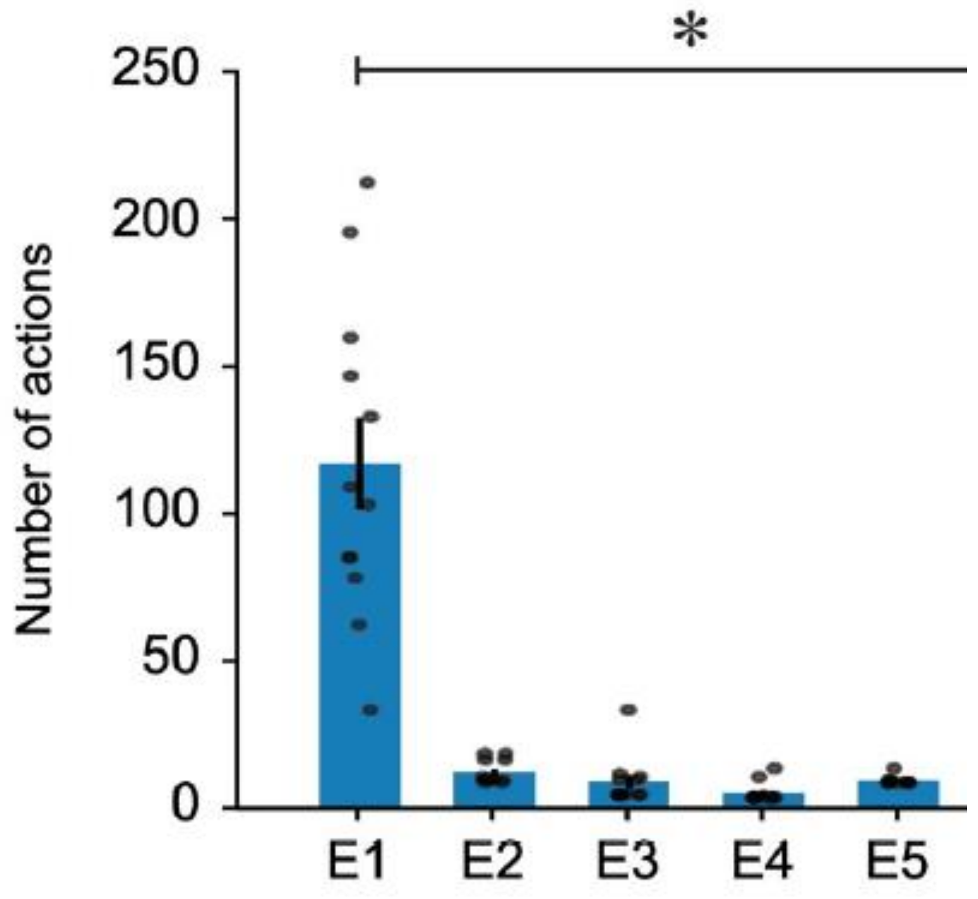
The role of exploration, novelty, and surprise in RL

1. Definitions of Novelty and Surprise (tabular environment)
2. Why is Surprise useful?
3. Change-point detection by Bayes-Factor Surprise
4. Why is Novelty useful?
5. Hybrid Model with Novelty, Surprise, and Reward
6. An Experiment

Environment: Markov Decision Process







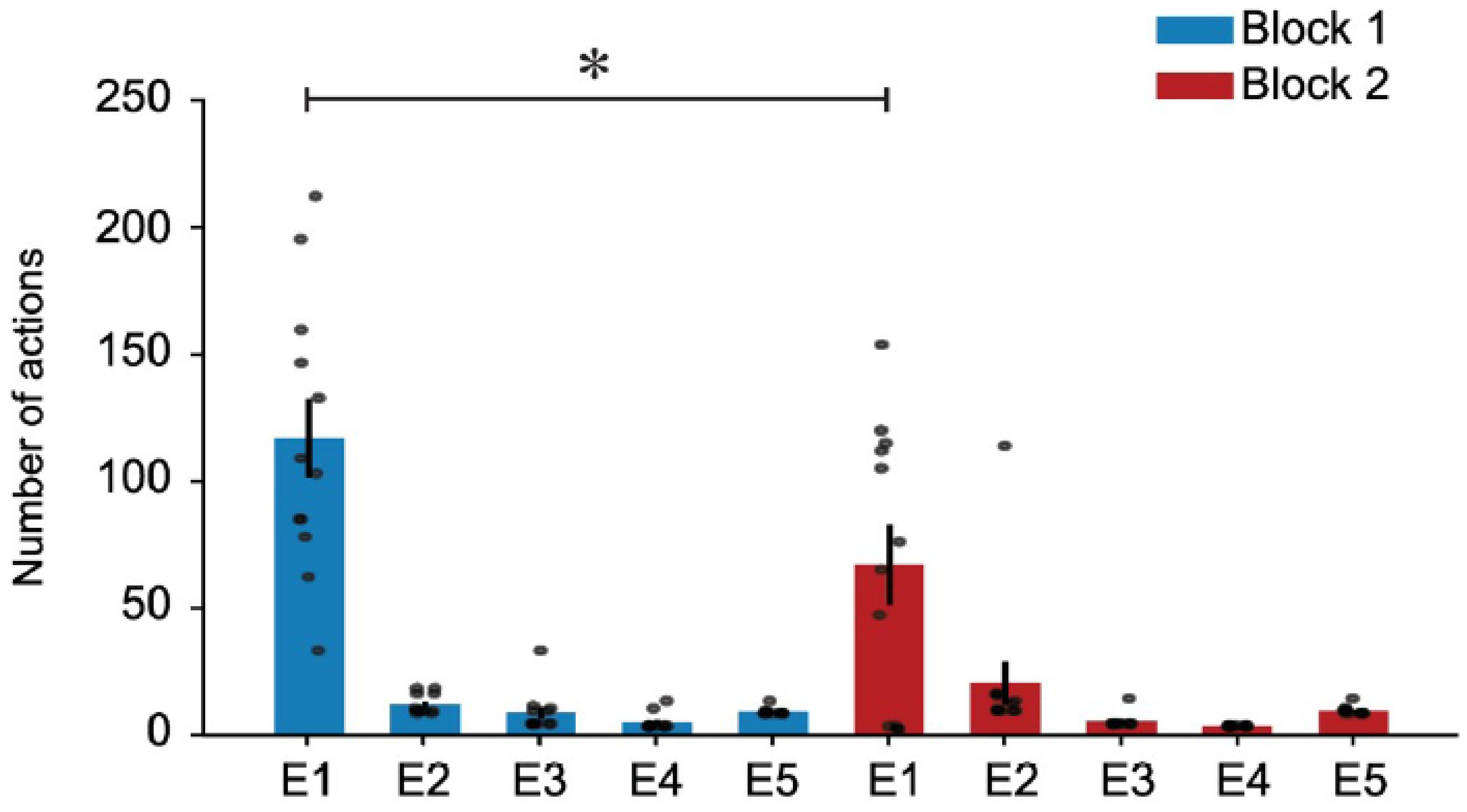
Finding 1)
Participants need about 150 actions in episode 1

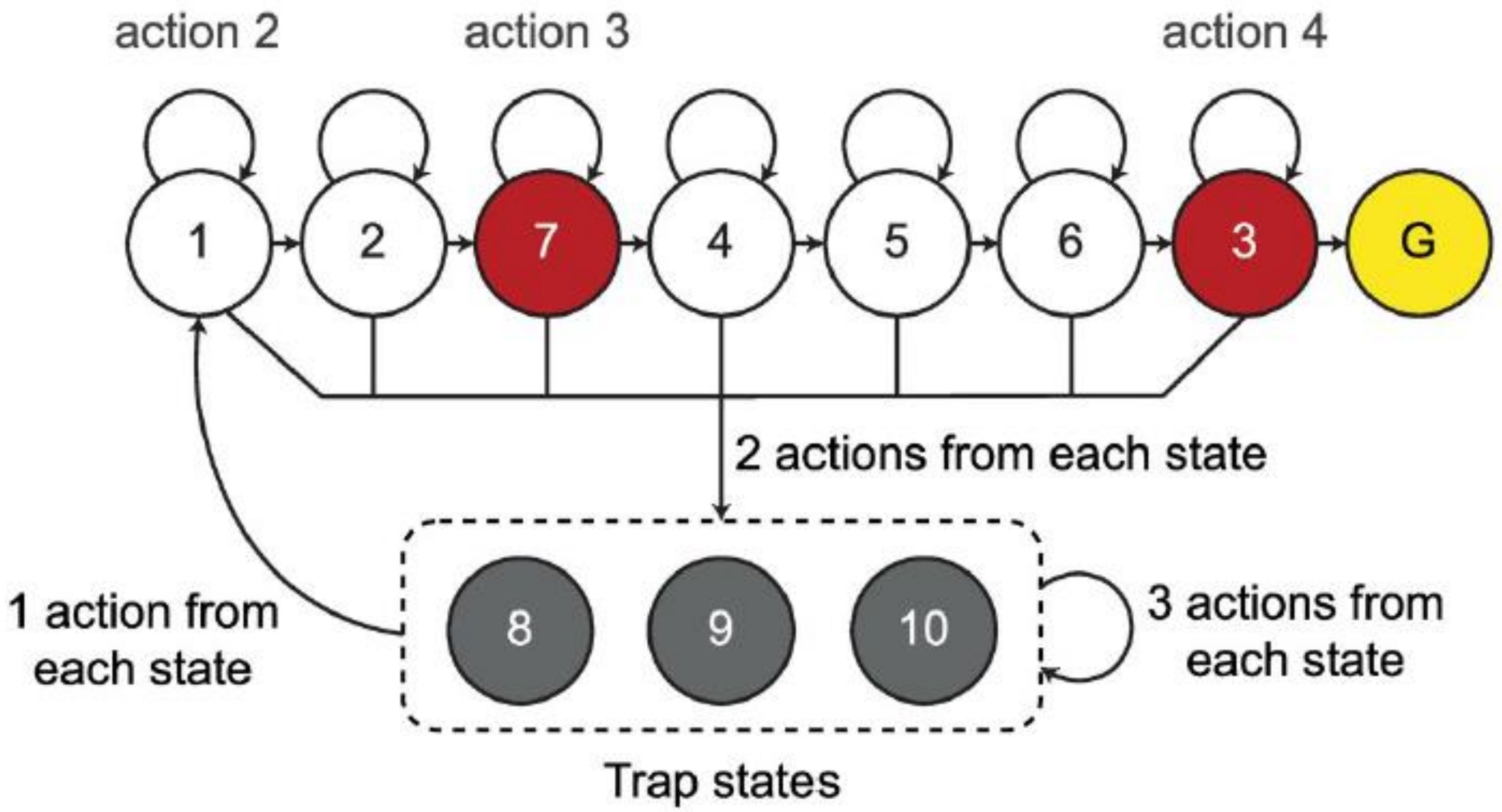
Finding 2)
In episode 2, participants go straight to goal

Volatile Environment: Switch after episode 5

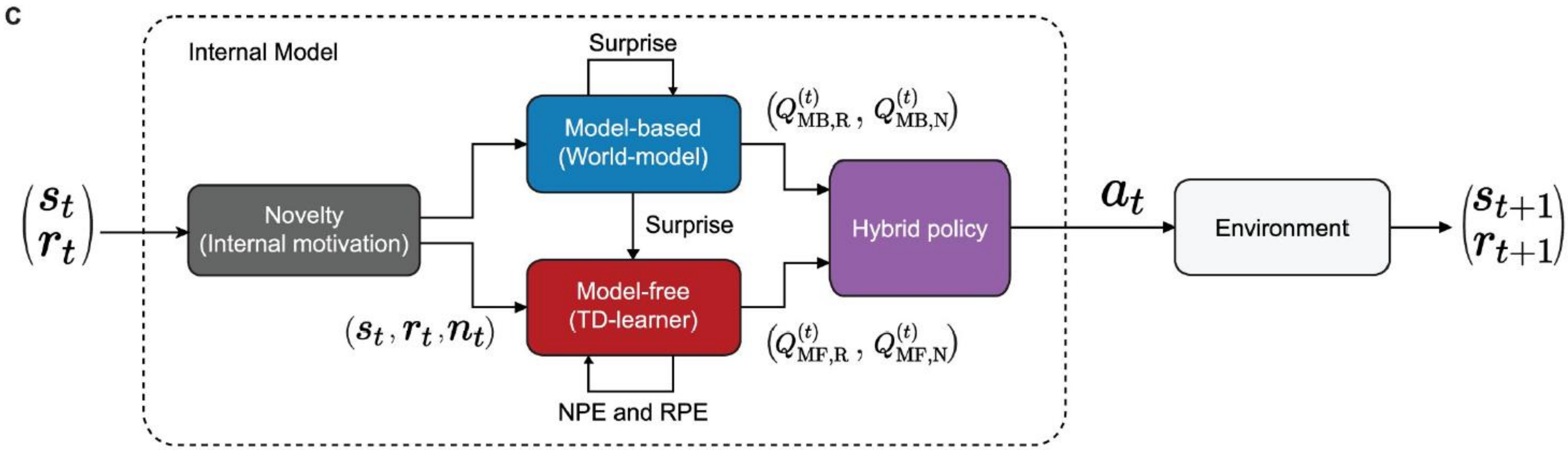
Finding 3)
In episodes 5 and 6, participants rapidly relearn!

Questions:
- Is Surprise necessary to explain relearning?
- Are humans model-based or model-free?
- Is novelty a good explanation of results?





Review: Hybrid model with separate paths Surprise, Novelty, Reward (SurNoR)

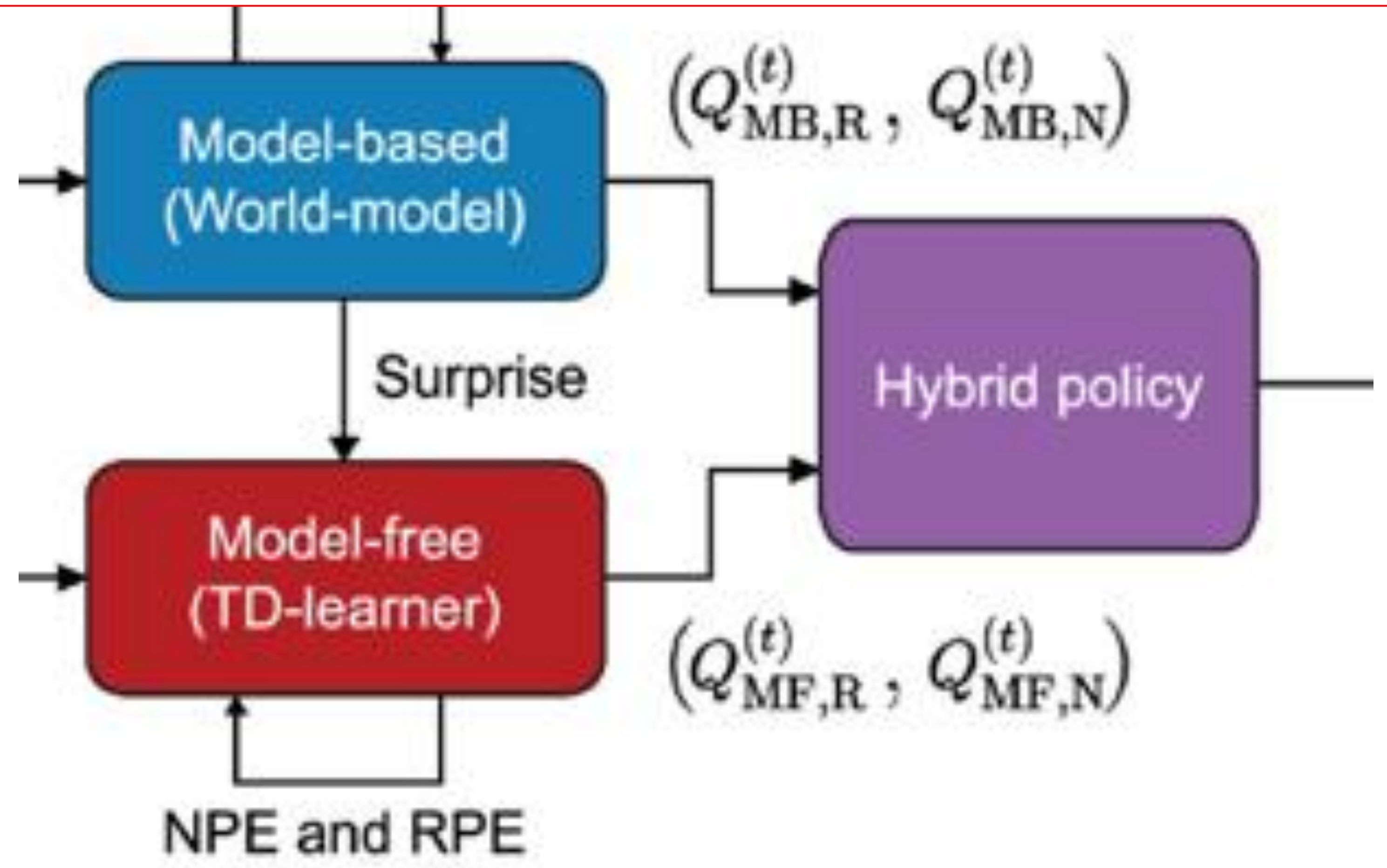


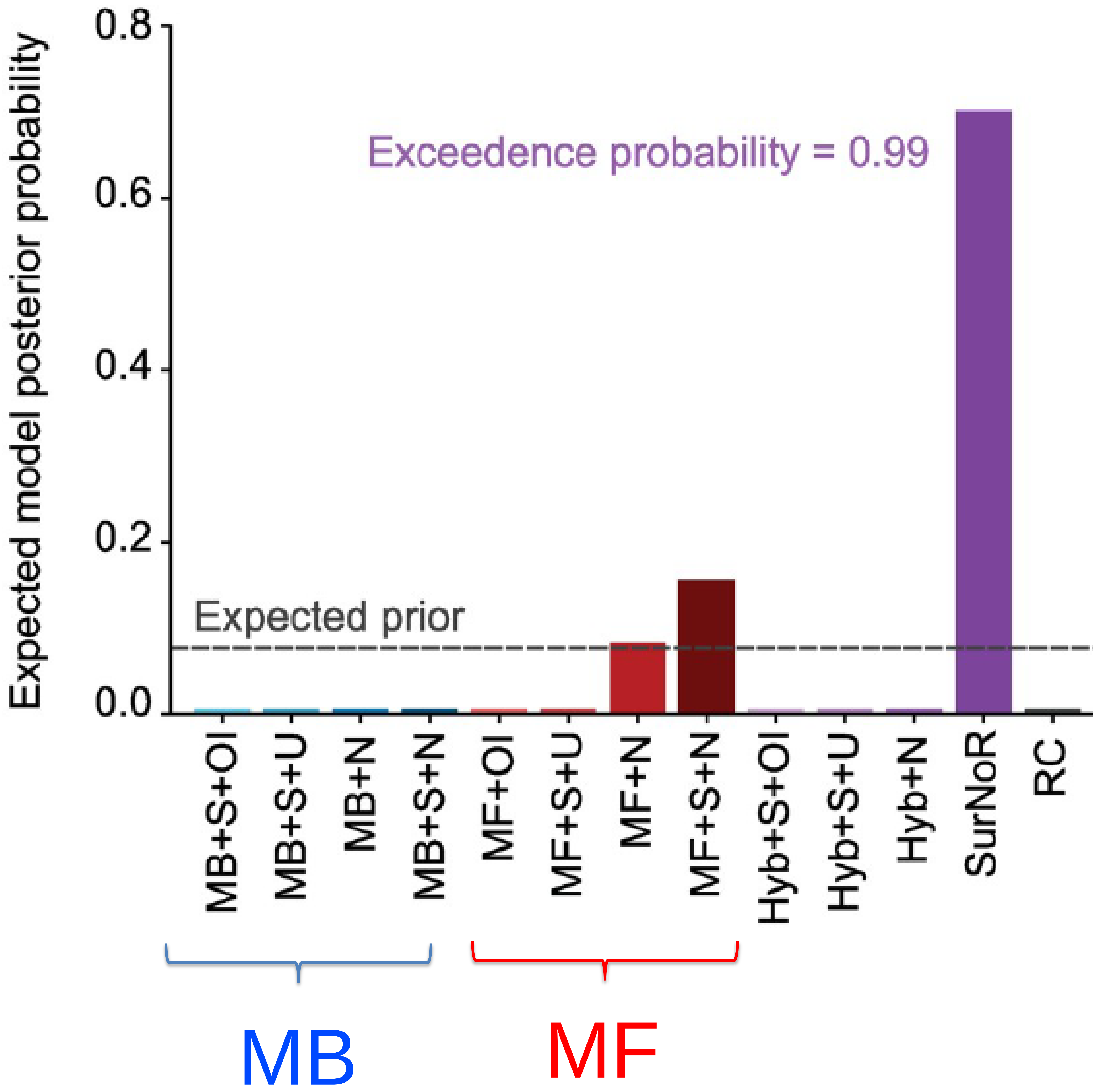
$$\text{RPE} = [r_t + \gamma \max_{a'} Q_R(s', a') - Q_R(s, a)]$$

$$\text{NPE} = [n_t + \gamma \max_{a'} Q_N(s', a') - Q_N(s, a)]$$

Comparison of Models: Surprise, Novelty, Reward

Finding 4)
Rapid relearning needs surprise



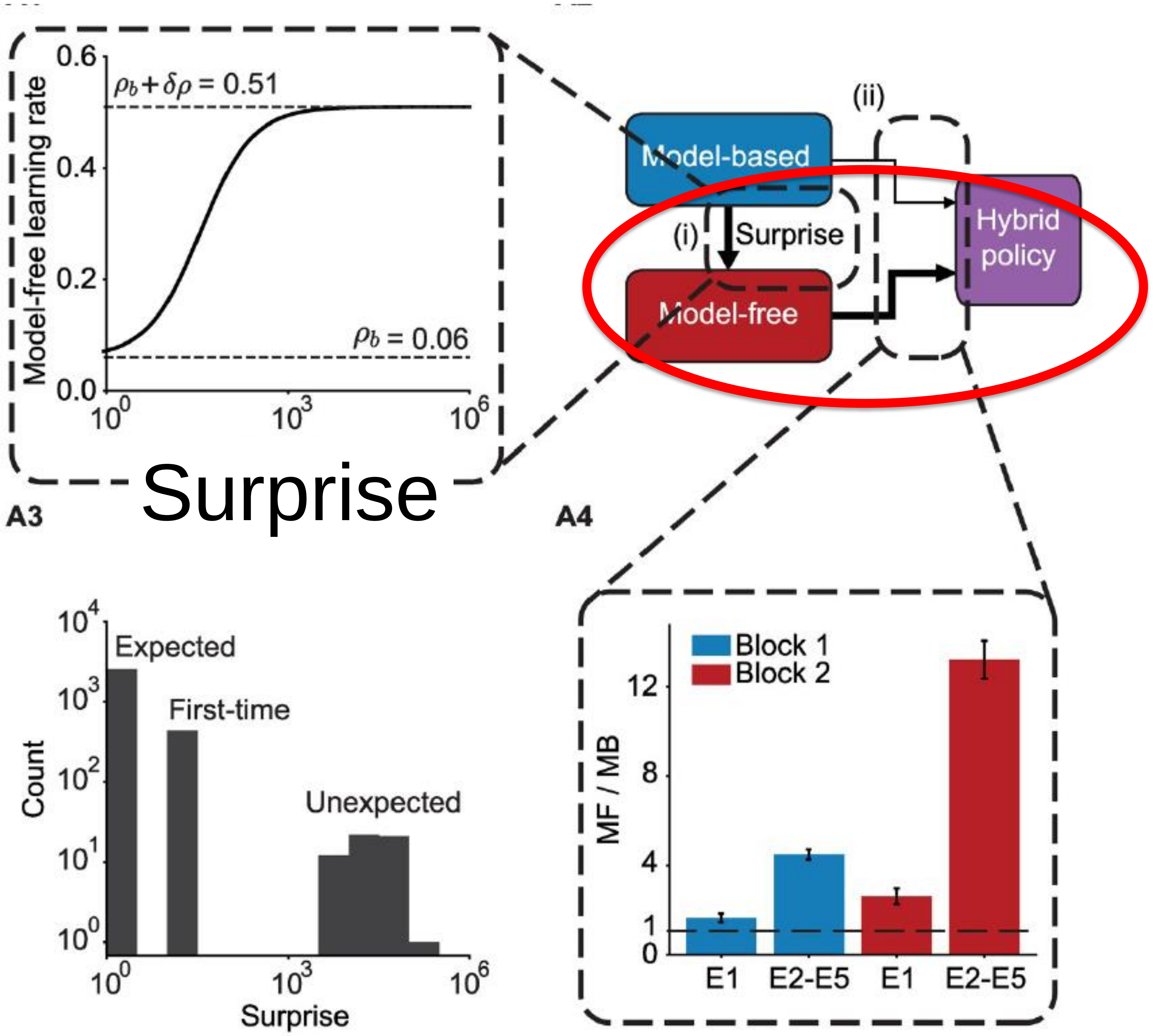


- Turn off novelty
- Turn off surprise
- Turn off model-based →MF
- Turn off model-free →MB
- OI = Optimistic Initialization

Relative importance of model-based versus model-free

Finding 5)
Model-free dominates
Human behavior!

surprise-modulated learning rate



Surprise is used modulate learning in RL

Finding 6)
Surprise is against expectations.
Hence surprise needs a **world model.**

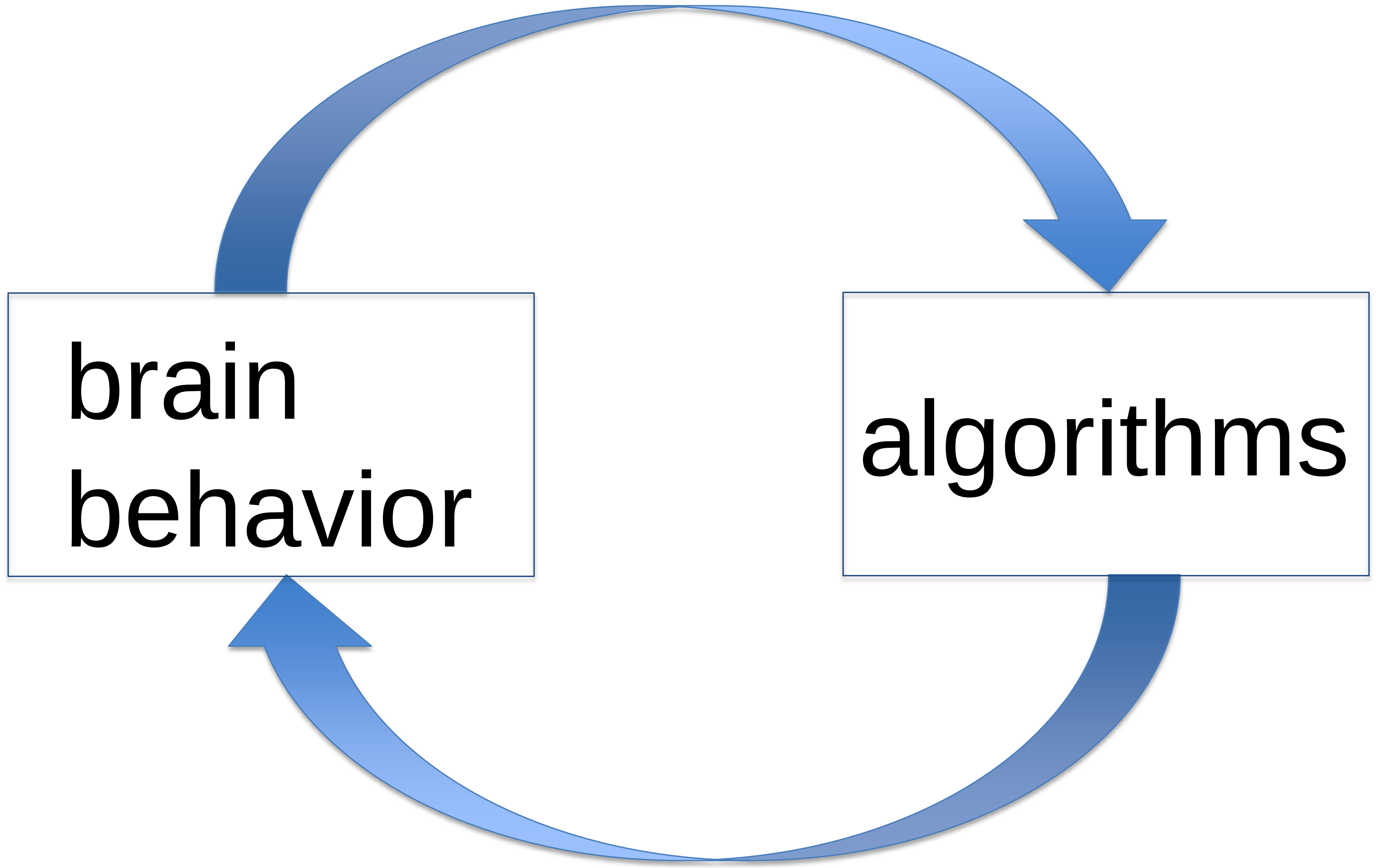
However, world model is
- Not used to do planning!
- Only used to extract surprise!

World-model not used for planning!

Reward-based learning versus Surprise-based learning

Reward-Prediction Error	→	Surprise
defined as TD error	→	defined as Bayes Factor Surprise
stimulated by chocolate, money, praise, …	→	stimulated by observations not consistent with momentary nodel of environment
modulates learning rate	→	modulates learning rate

Current Research in Reinforcement Learning:



- Exploration → not exploration bonus, but separate modules
- Novelty → Novelty supports exploration
- Surprise → Surprise detects changes/adapts learning

Thanks!

The END

… of part 1 for today.
We talk about exam procedures next week.