UNIVERSITÉ TOULOUSE 1 CAPITOLE

IRIT
CNRS - INPT - UPS - UT1 - UTM

LILaC
Logique, Interaction, Langue, et Calcul

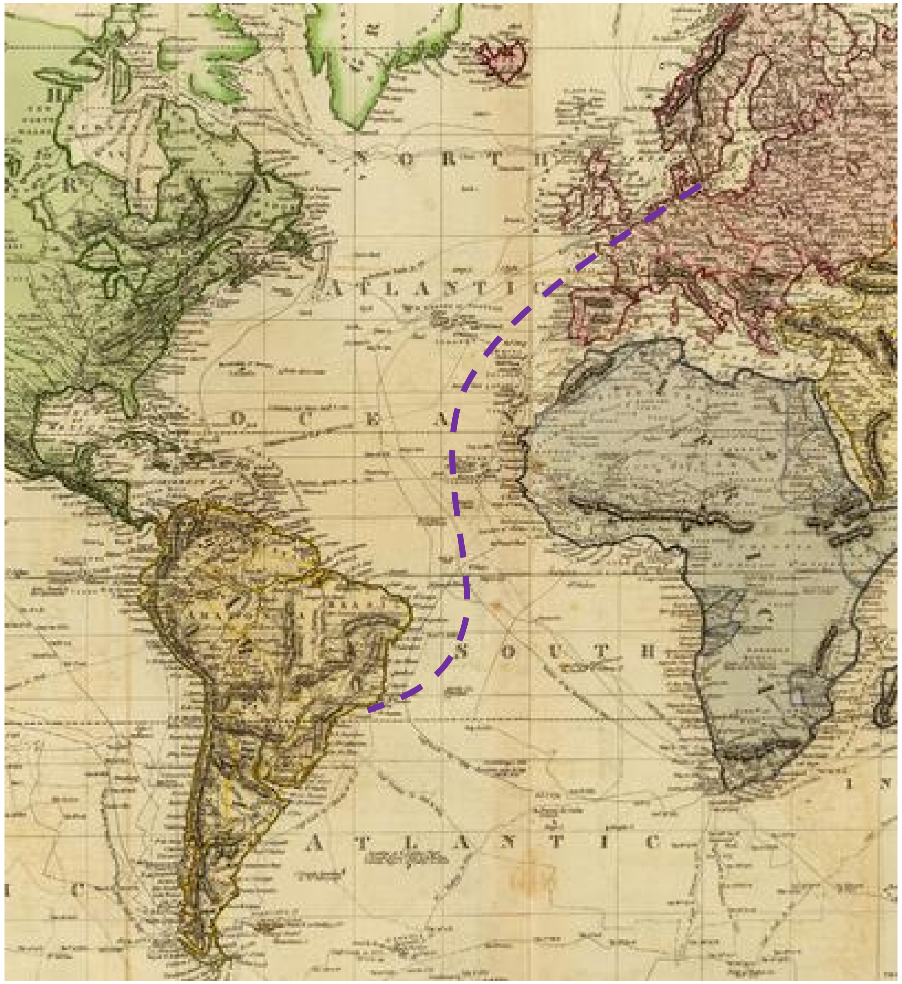
Looking for the right time to shift strategy in the Exploration-Exploitation Dilemma

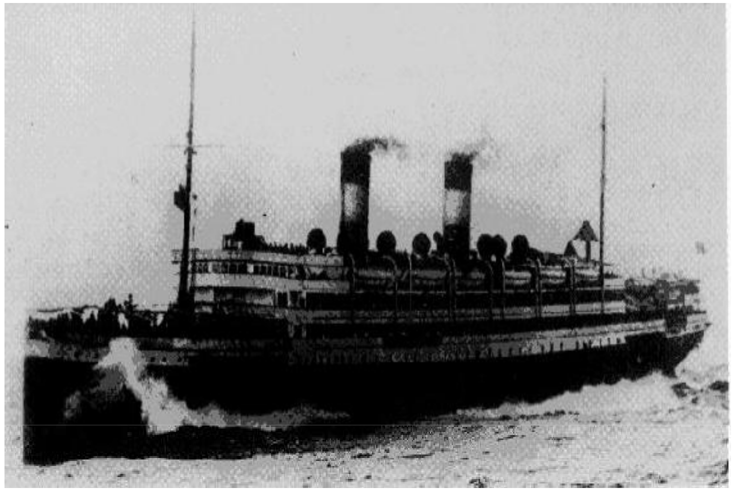
Filipo STUDZINSKI PEROTTO

tfml 2015
theoretical foundations of machine learning, Będlewo

2015 February

Anecdote





XIX Century
Immigrants

Summary

- Exploitation-Exploration Dilemma
- Generalities
- Standard Approaches x Proposed Approach
- Experimental Setting and Results
- Conclusions and Perspectives

Exploration – Exploitation Dilemma

Exploration-Exploitation Dilemma

Should I…

…try something new?   or just…   enjoy a known pleasure?

Exploration-Exploitation Dilemma

Any kind of adaptive agent...

(humans, animals, robots, societies, companies, ...)

...must perform and learn at the same time.

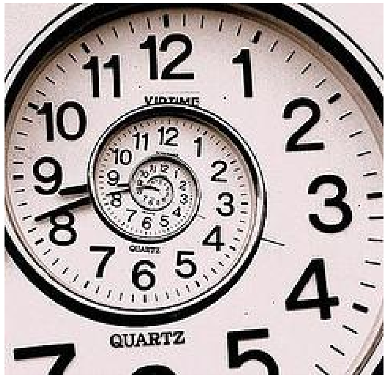
– how to balance it?

– more knowledge = better performance

– but learning = investment

Exploration-Exploitation Dilemma

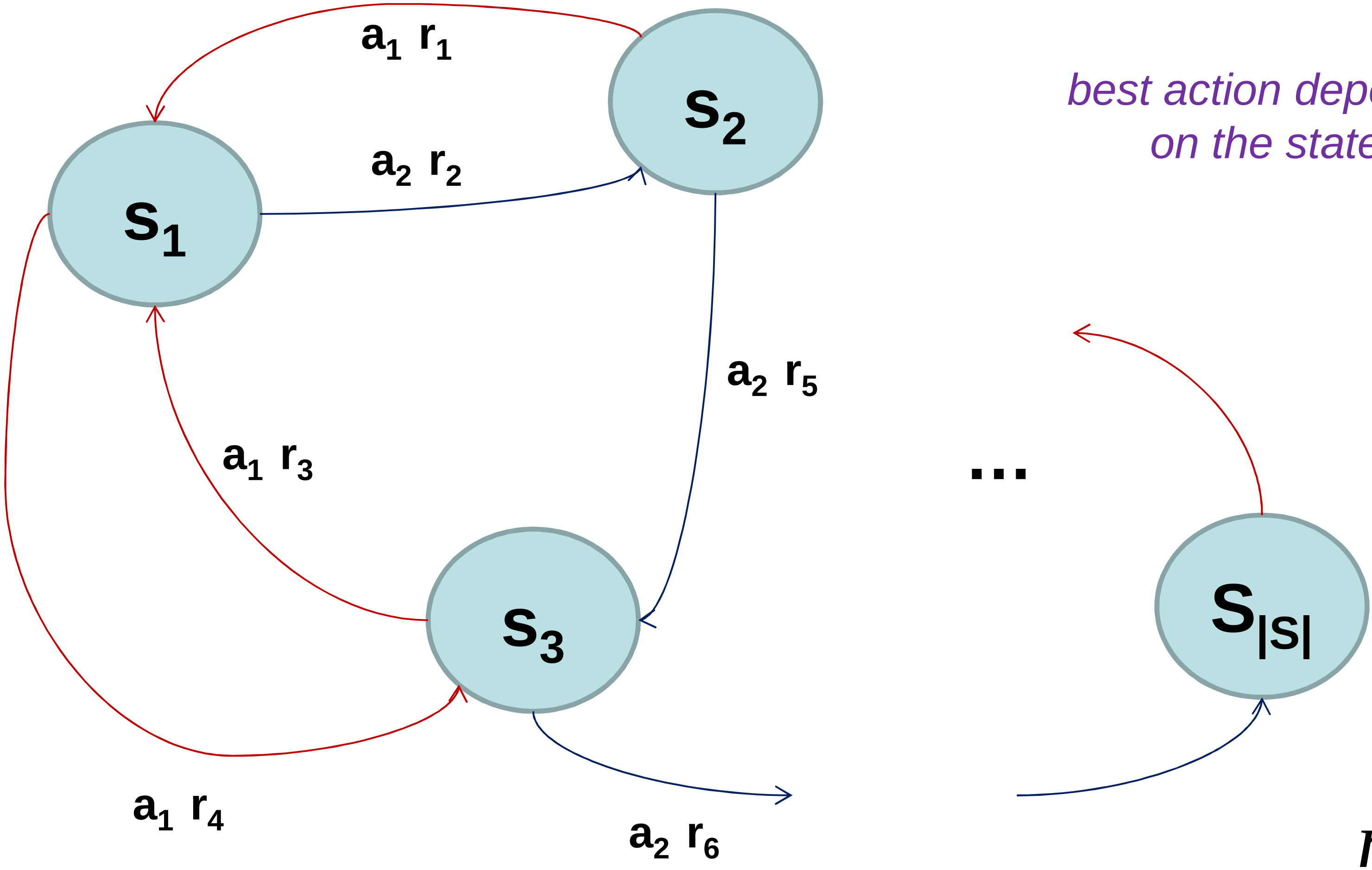
- When to explore ?



- How to explore ?

Generalities

Markovian Decision Process



best action depends on the state

$r \in \Re$

Markovian Decision Process (MDP)

$\boldsymbol{S} = \{s_1, s_2 \dots s_{|S|}\}$ set of states

$\boldsymbol{A} = \{a_1, a_2 \dots a_{|A|}\}$ set of actions

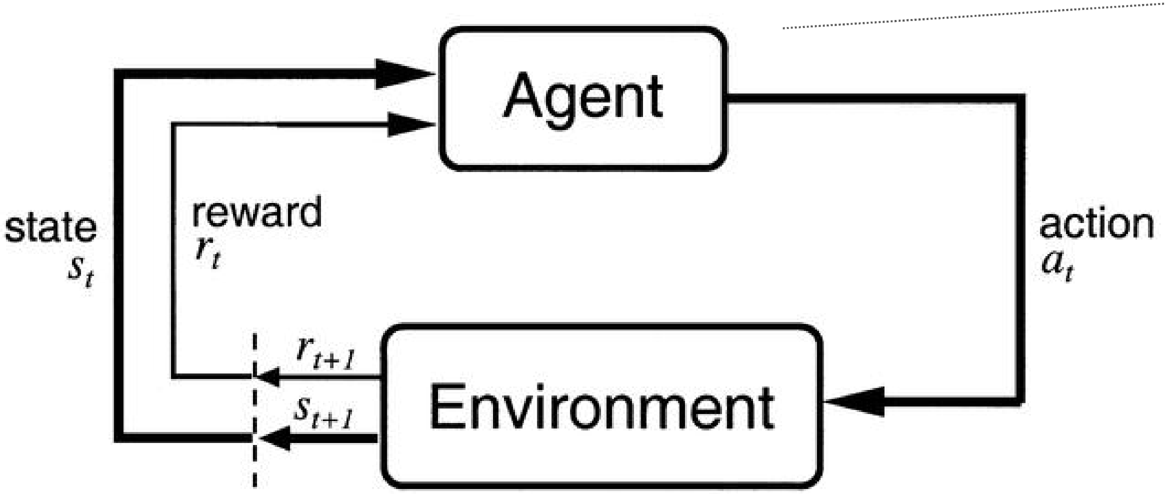
$\boldsymbol{T} = \Pr(s' \mid s, a)$ transition function

$\boldsymbol{R} = \Pr(r \mid s, a, s')$ reward function

Solution : a policy of actions ® $: S \rightarrow A$

which maximizes expected future rewards

Reinforcement Learning



- *Observe* ***s***
- ***Decide***
- *Execute* ***a***
- *Receive* ***r***
- *Observe* ***s'***
- ***Learn***

Non-episodic setting : continuous lifetime experience

Agent Lifecycle

- *Observe **s***
- ***Decide***
  - *Choose Strategy*
  - *Select Action (from policy)*
- *Execute **a***
- *Receive **r***
- *Observe **s'***
- ***Learn***
  - *Update Models*
  - *Calculate Utility*
  - *Redefine Policy*

*model-based learning

Agent Lifecycle

- *Observe* ***s***
- ***Decide***
  - *Choose Strategy*
  - *Select Action (from policy)*
- *Execute* ***a***
- *Receive* ***r***
- *Observe* ***s'***
- ***Learn***
  - *Update Models*
  - *Calculate Utility*
  - *Redefine Policy*

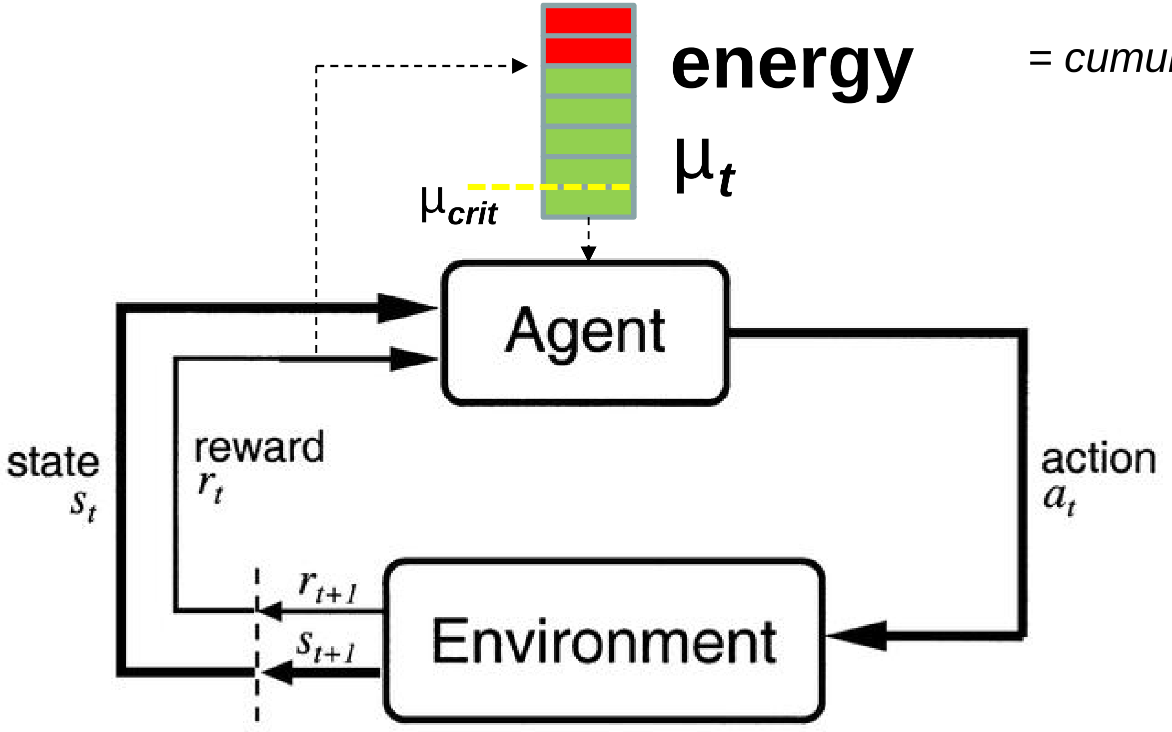
*model-based learning

Issues

RL with Limited Resources

- How about costs?

Reinforcement Learning with Energy



= cumulated reward + initial energy

$$= \omega_0 + \sum r_t$$

New challenge : managing costs, alert under ω_{min}

Standard Approaches

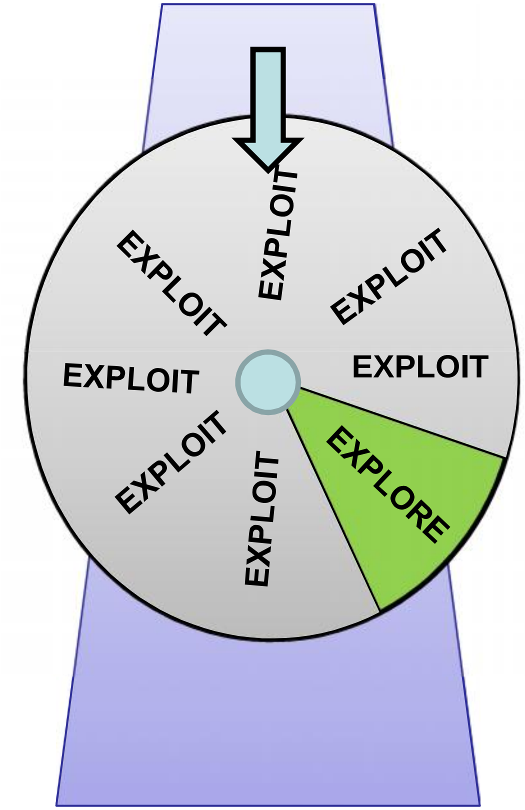
Epsilon-Greedy Method

- Choose exploration

  with probability £

- Choose exploitation

  with probability 1-£

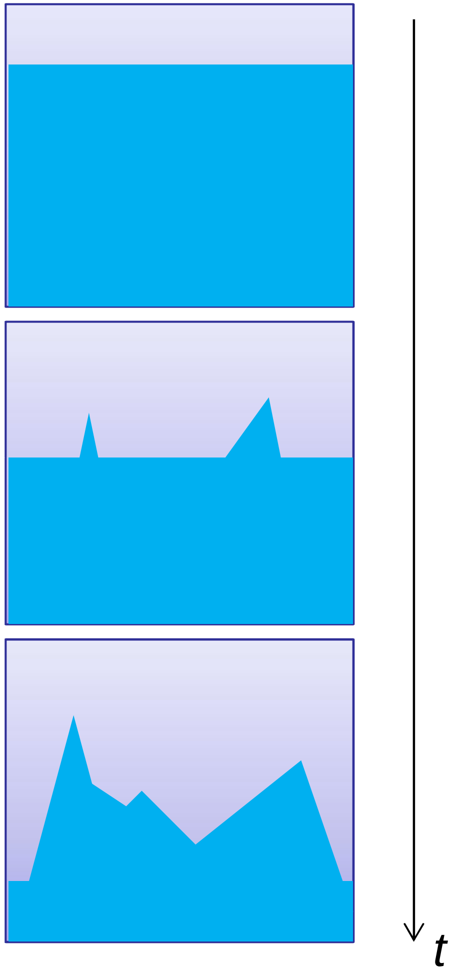


* $\{\varepsilon \in \Re \mid 0 \leq \varepsilon \leq 1\}$

Optimism in face of incertitude

- Optimistic initialization
  - Optimistic-greedy
  - R-max
    - $Q_0(s, a) > Q^*(s, a)$
    - $\hat{R}_0(s) > R(s)$

- Exploration bonus
  - UCB
    - Uncertain or infrequent states



How to explore ?

- <u>Undirected</u> :

  = Random

  "when exploring,
  do something unexpected"

  $\rightarrow$ choose action at random

- <u>Directed</u> :

  = Exploration Bonus

  "when exploring,
  go towards unknown situations"

  $\rightarrow$ search less visited states
  $\rightarrow$ or more unpredictable states

Standard Approaches Difficulties

- Random Exploration
  - not very efficient in general
  - not like curiosity
  - can leads the agent to known bad choices

- Epsilon Methods
  - not efficient for sequential problems

- Optimistic Methods
  - efficient exploration
  - but long forced initial training time
  - costs not taken in account

Proposed Approach

Engaged Climber

- Intuition :
  - Choose the strategy (explore / exploit)
  - Stay engaged for a while (until to reach a “peak”)
  - Except in critical situations

Engaged Climber

- 2 Policies :
  - π_R : policy for exploiting
  - π_K : policy for exploring

- 2 Utilities:
  - V_R , Q_R : Exploiting Utility based on reward
  - V_K , Q_K : Exploring Utility based on uncertainty

Engaged Climber

- Parameters :

  ξ : time in the current strategy

  ξ_{max} : maximum engagement time

  r_{peak} : peak reward

  ω_{crit} : critical energy level alert

  ε : exploration rate

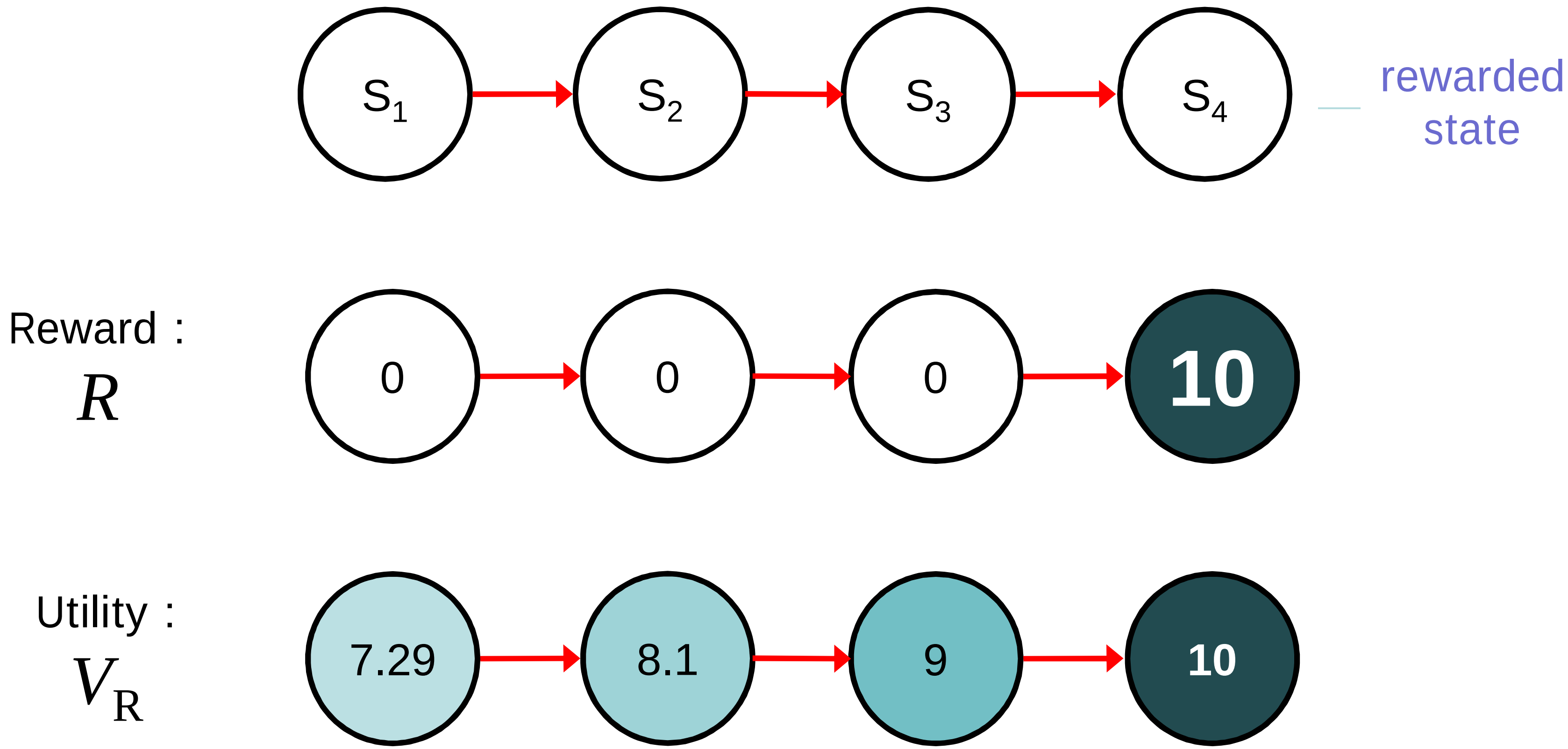
Engaged Climber

- Update Utilities :

$$\underbrace{Q_R(s, a)}_{\text{Reward Utility}} \quad \text{ê} \quad \underbrace{R(s, a)}_{\text{Immediate Reward}} + \underbrace{\gamma}_{\text{Discount Factor}} . \underbrace{\Sigma[T(s, a, s').V_R(s')]}_{\text{Expected Future Reward Utility}}$$

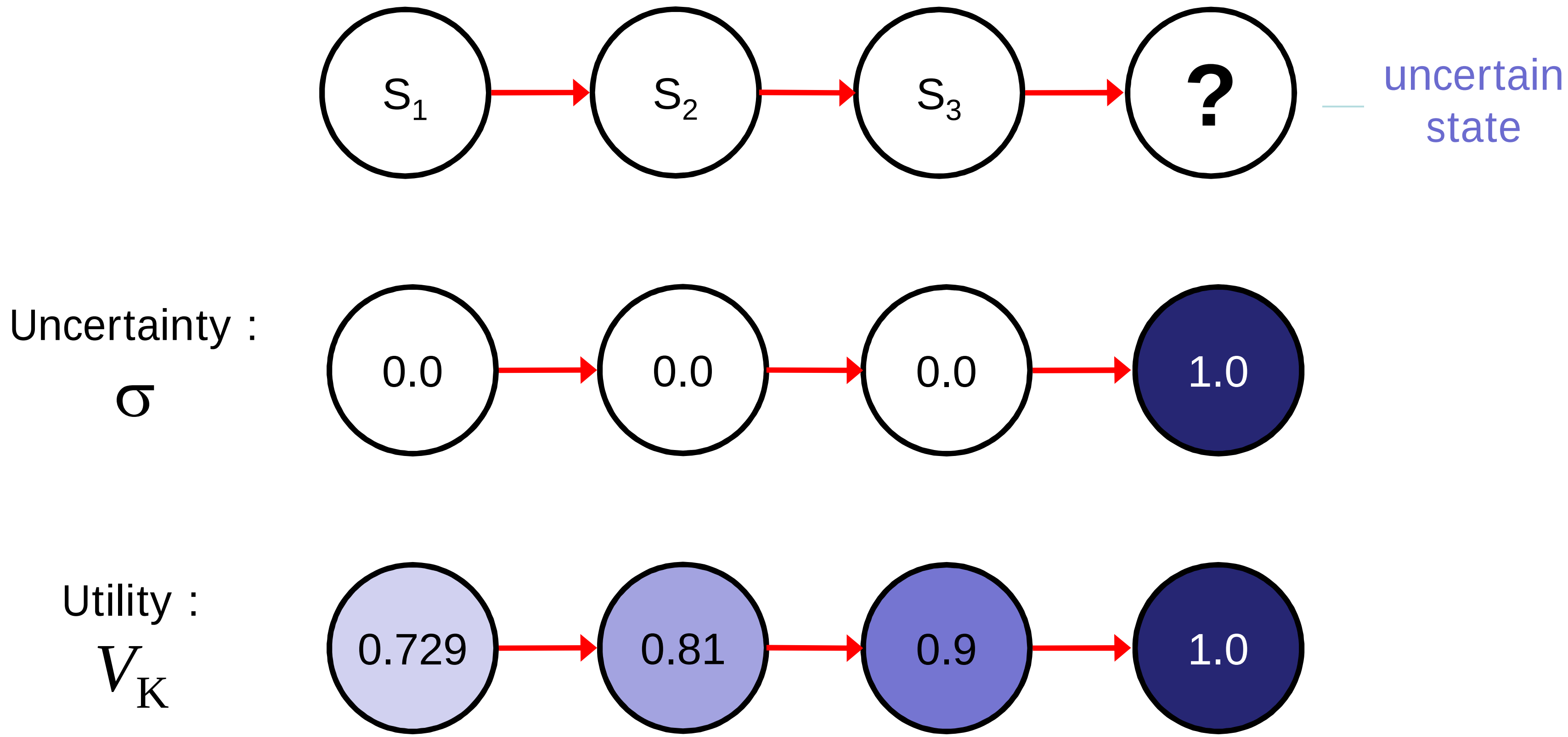
$$\underbrace{Q_K(s, a)}_{\text{Learning Utility}} \quad \text{ê} \quad \underbrace{\pm(s, a)}_{\text{Immediate Uncertainty}} + \underbrace{\gamma}_{\text{Discount Factor}} . \underbrace{\Sigma[T(s, a, s').V_K(s')]}_{\text{Expected Future Learning Utility}}$$

Reward Utility



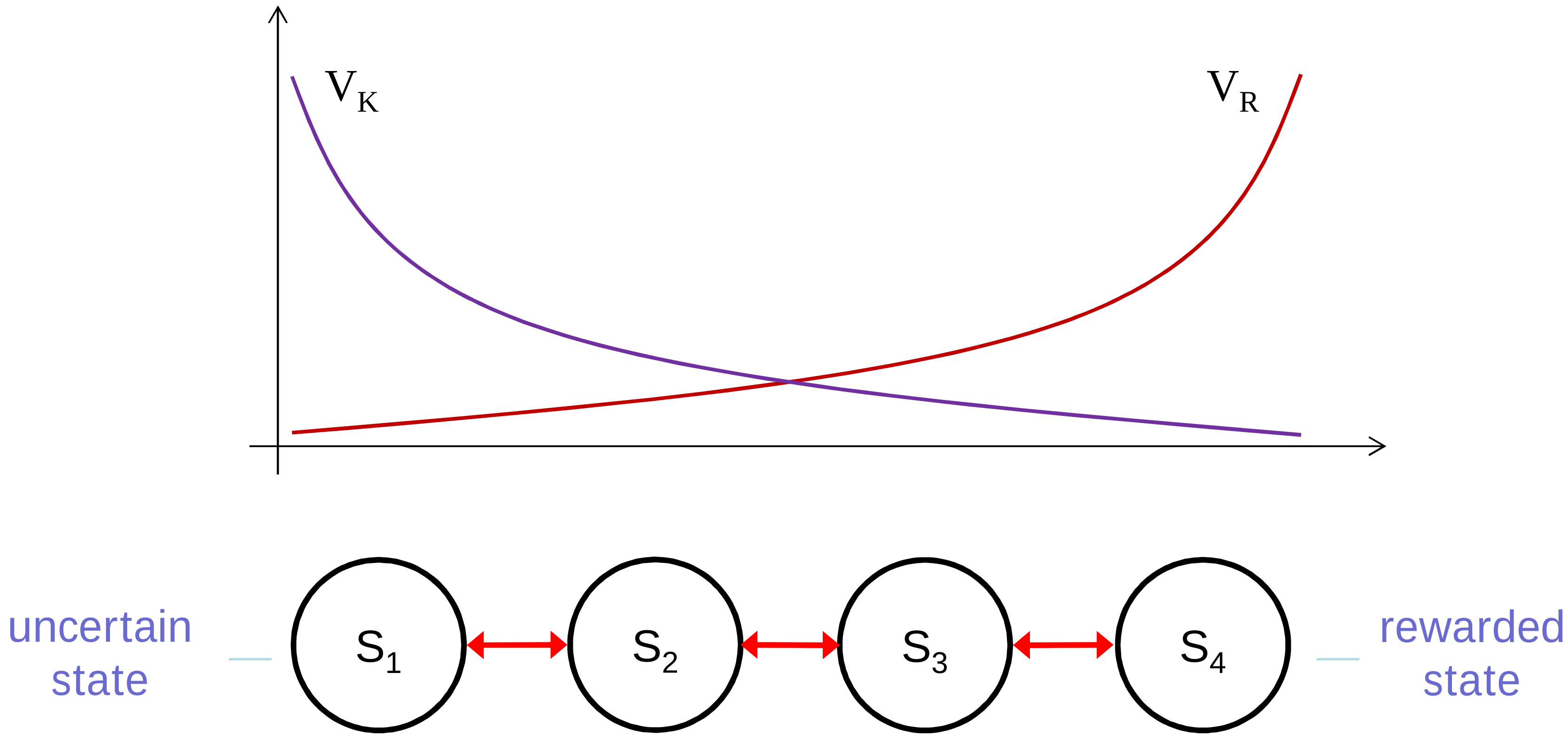
discount factor γ = 0.9

Learning Utility

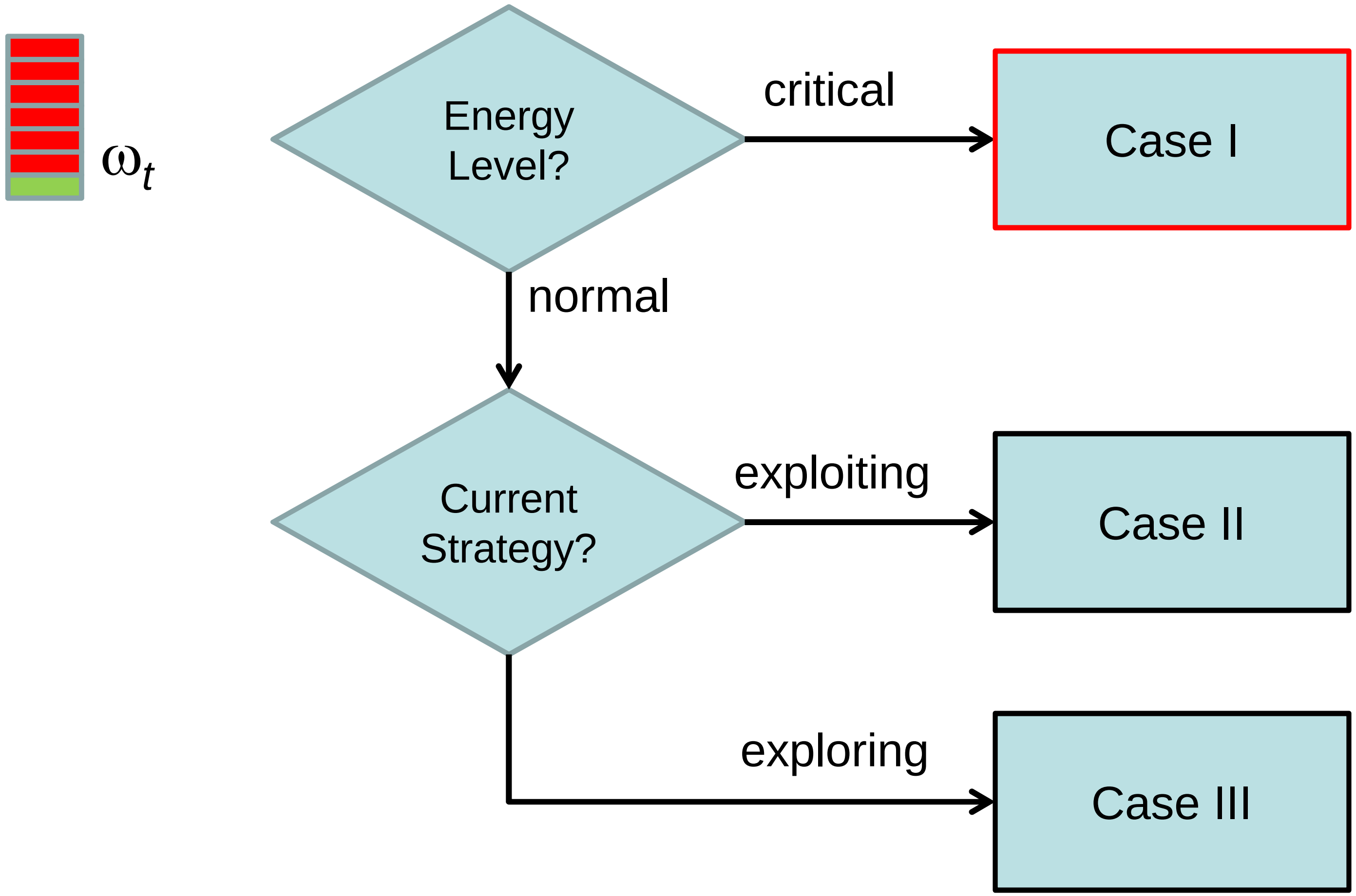


discount factor γ = 0.9

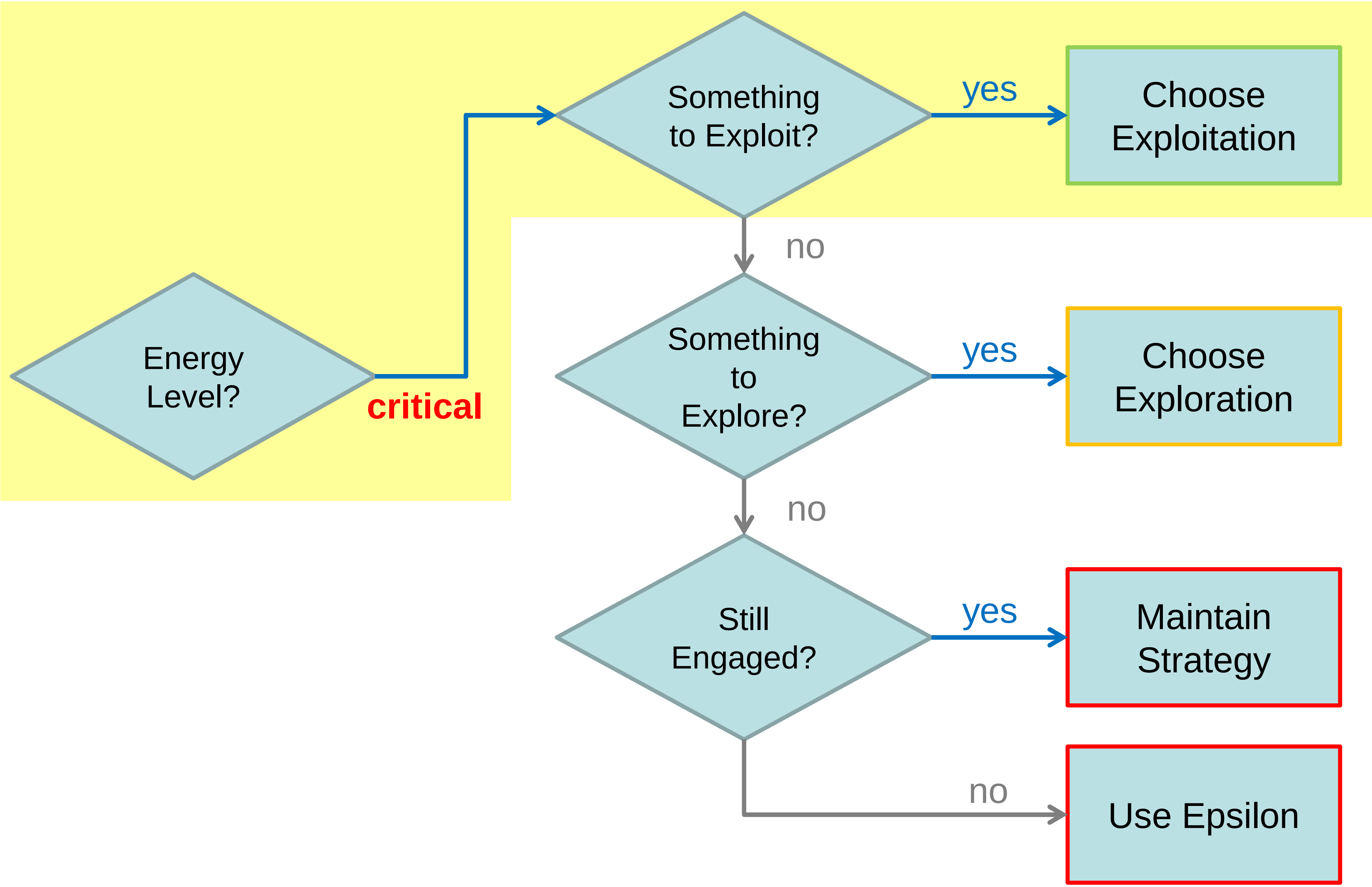
Utilities



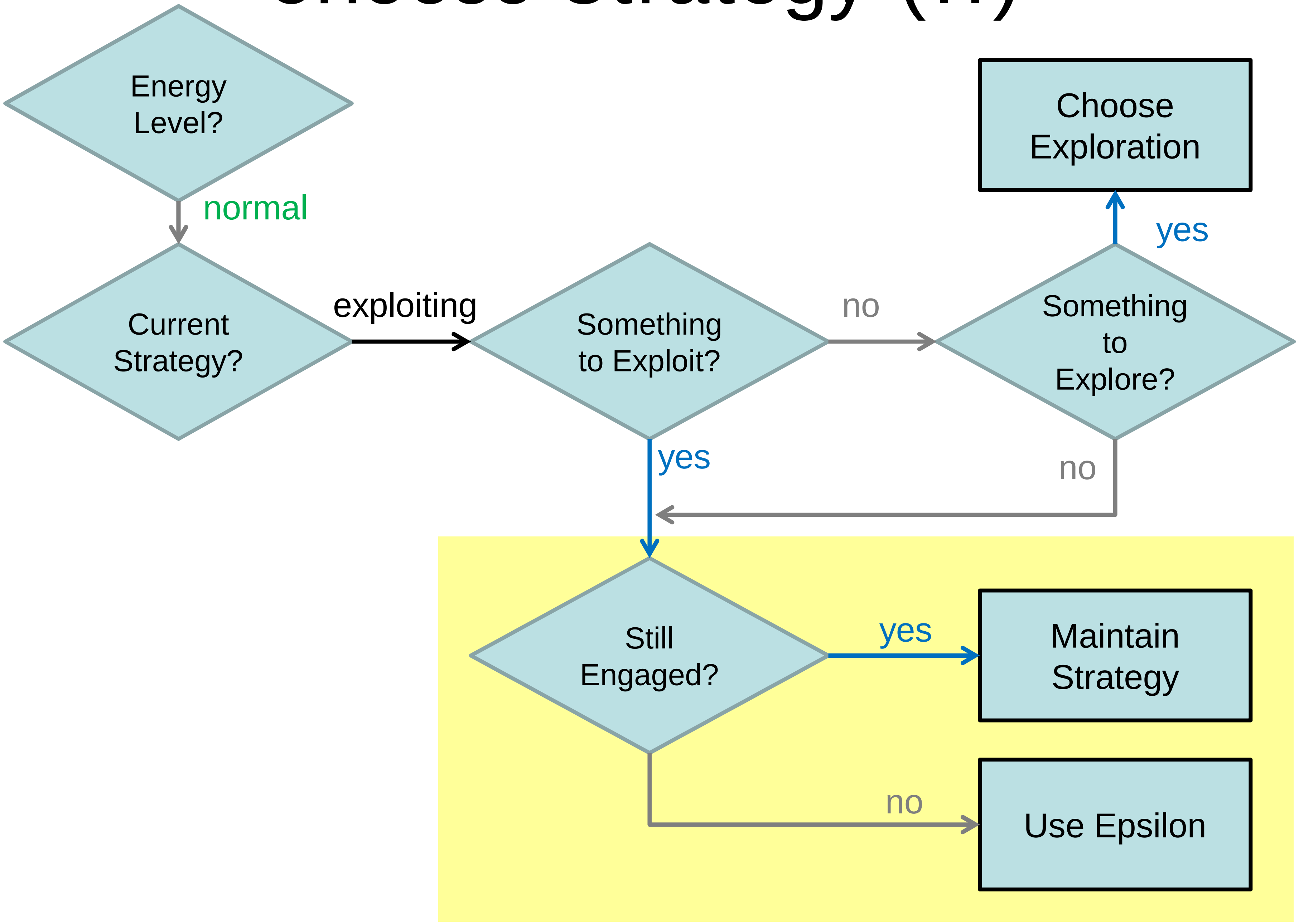
Choose Strategy (Engaged-Climber)



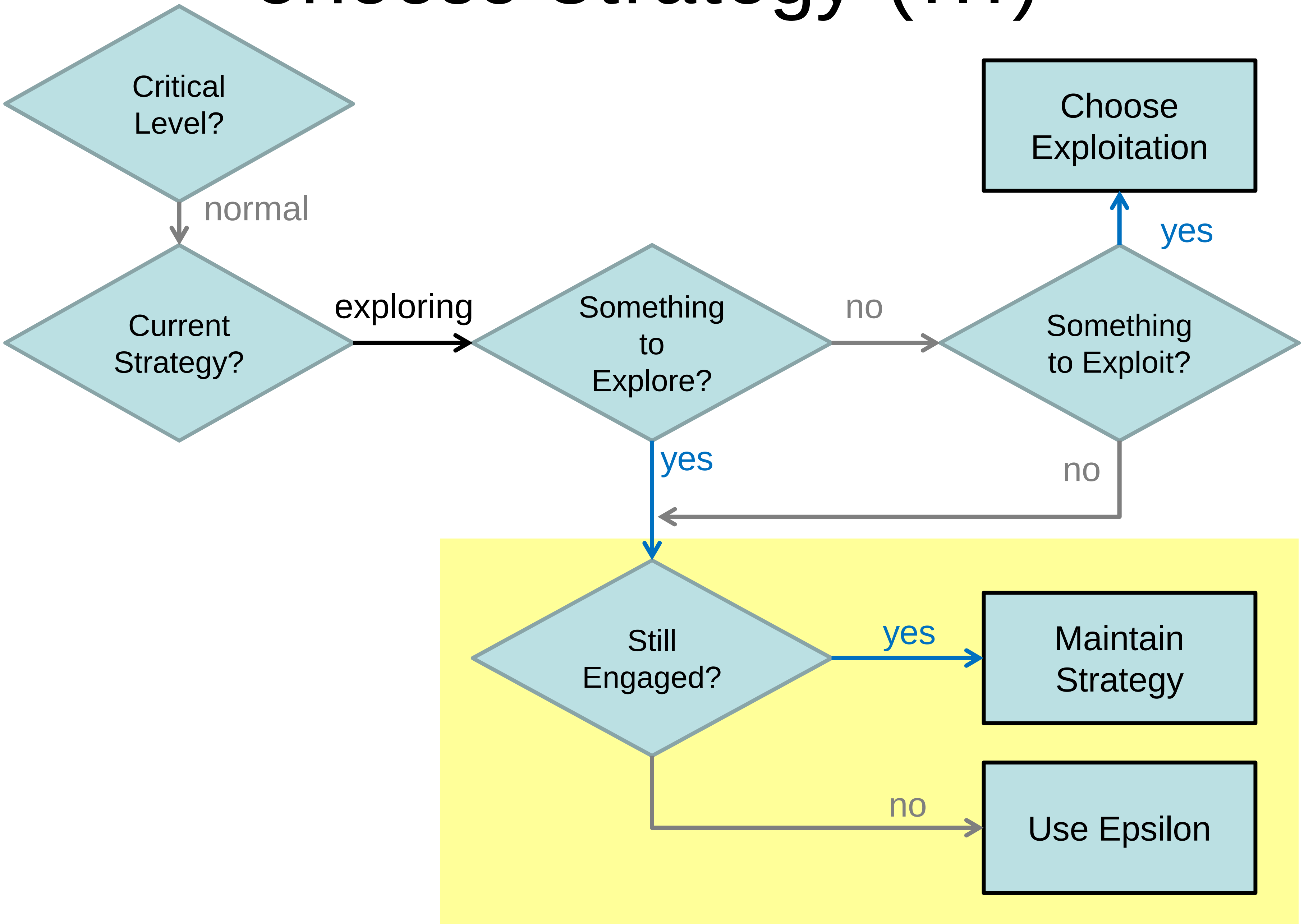
Choose Strategy (I)



Choose Strategy (II)

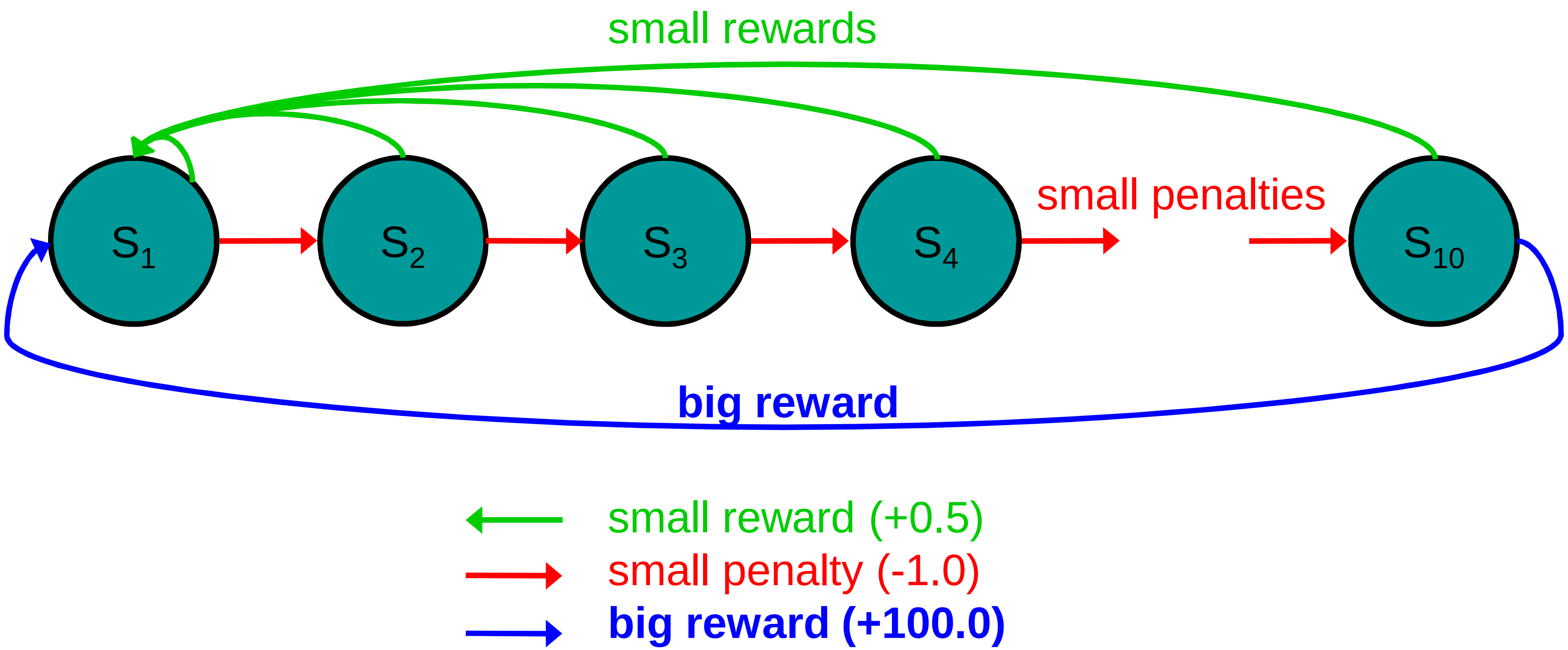


Choose Strategy (III)



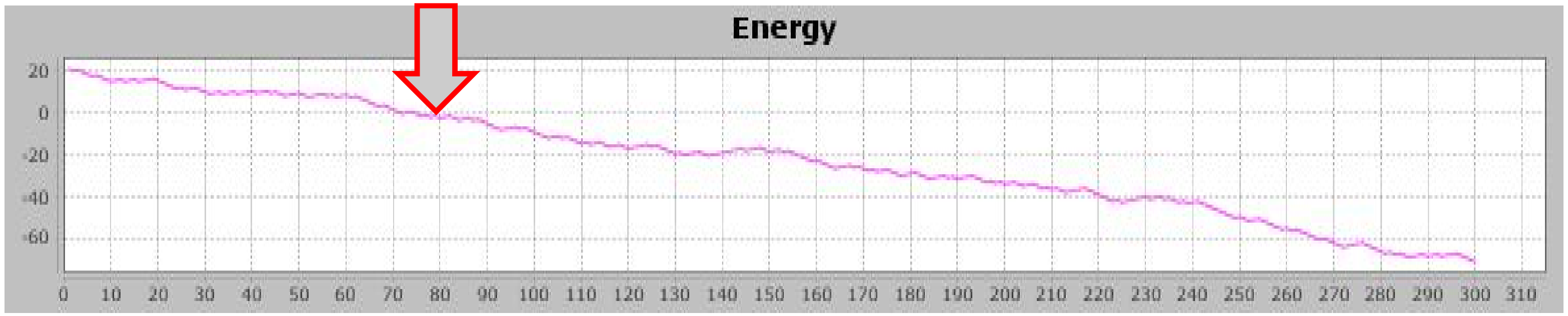
Experience : Chain States

- 10 states, 2 actions :
  - go forward
  - back to start

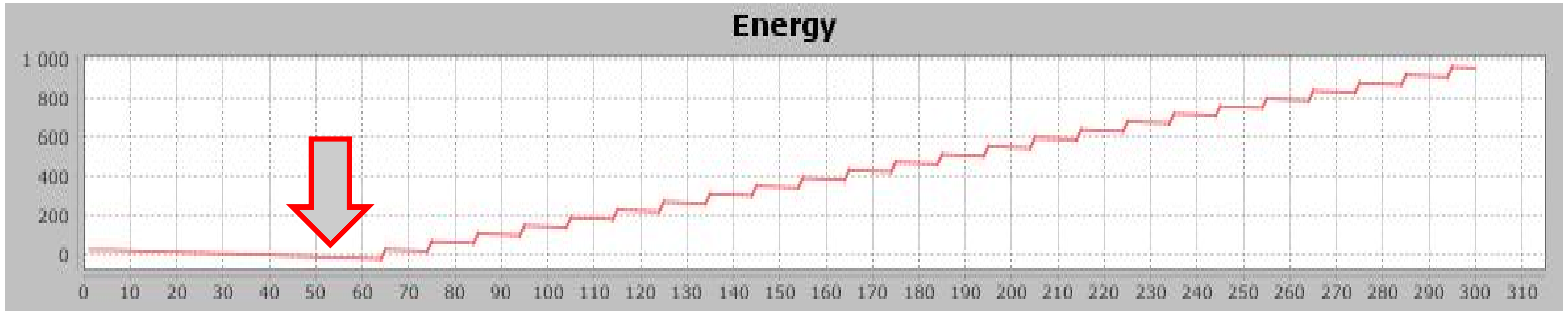


Experimental Results

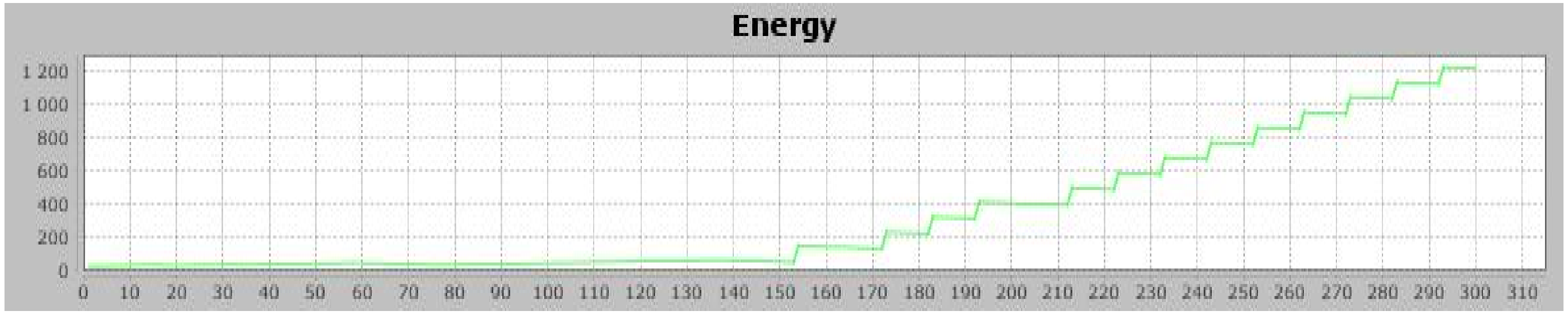
Eps



Opt



EC



Experimental Results

Table 1.: Experimental results

Method	Time to discover the goal	Minimum Score
ε-Greedy	≈ 1200	≈ -400
Optimistic-Greedy	≈ 60	≈ -30
Engaged-Climber	≈ 170	≈ +10

**using value-iteration method for calculating utilities*

Conclusions

- Curiosity
  - Policy for Exploring
  - Policy for Exploiting
- Engagement
  - More human-like behavior
- Consider limited resources

Perspectives

- Concrete search for peaks
- Factored Representations
- Partial Observation
- Multiple Policies
- Robustness

UNIVERSITÉ TOULOUSE 1 CAPITOLE

IRIT
CNRS - INPT - UPS - UT1 - UTM

LILaC
Logique, Interaction, Langue, et Calcul

Looking for the right time to shift strategy in the Exploration-Exploitation Dilemma

Filipo STUDZINSKI PEROTTO

tfml 2015
theoretical foundations of machine learning, Będlewo

2015 February