SMART School on Computational Social and Behavioral Sciences

Reinforcement learning in animals, from the standpoint of navigations

Benoît Girard

benoit.girard@isir.upmc.fr
Institut des Systèmes Intelligents et de Robotique (ISIR)

September 2017

Plan

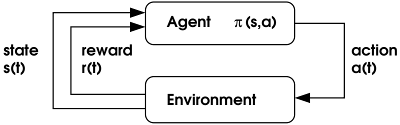
1. **Introduction**
   - Goal
   - Model-based & Model-free RL
   - Neural substrate of Navigation
2. **Navigation strategies**
   - Taxonomies
   - Navigation strategies: what & how?
3. **Multiple system interactions**
   - (Dollé et al., 2010)
   - (Caluwaerts et al., 2012a,b)
4. **Conclusion**

Plan

Multiple reinforcement learning algorithms / behavioral strategies / navigation methods

- Reinforcement learning, as formalized in AI:
  - has been quite successful at explaining animal behavior in instrumental conditioning,
  - has interesting links with the physiology of dopamine.
- Different families of algorithms predict different adaptation patterns to changes.
- This is quite obvious in navigation tasks, where multiple strategies are used by animals.
- But navigation also invites us to investigate:
  - how multiple RL systems can collaborate,
  - behavioral systems beyond RL.

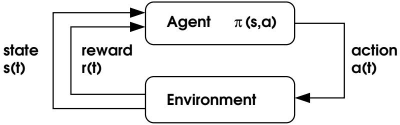
Reinforcement Learning formalism



Unsupervised learning

- occasional reward/punishment feedback,
- no precise information about the changes to be made,
- long sequences can cause the reinforcement feedback: *temporal credit assignment problem*
- Numerous algorithms (Sutton & Barto, 1998).

Reinforcement Learning formalism



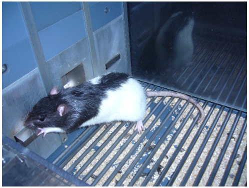
Goal

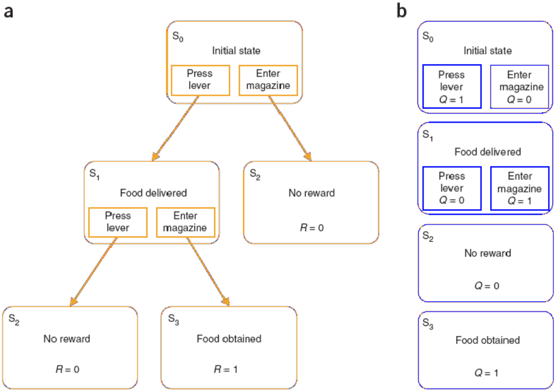
Find the policy $\pi(s, a)$ maximizing the return R.

Often formalized as:

$$R_t = r_{t+1} + \gamma r_{t+2} + \gamma^2 r_{t+3} + \ldots = \sum_{k=0}^{\infty} \gamma^k r_{t+k+1}, \text{ with } 0 < \gamma < 1$$

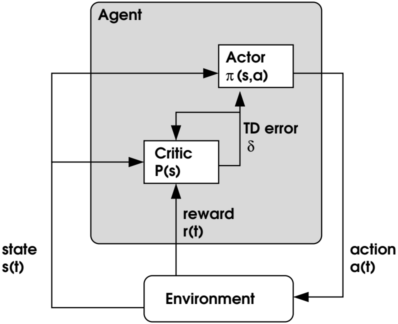
Model-based & Model-free learning algorithms





(Daw et al., 2005, Nat. Neurosci.)

Model-free RL (Actor/Critic example)



An example of temporal-difference (TD) learning algorithms.
Sutton's PhD thesis (1984) :

- The Critic learns to predict the value P_t of each state, so that $P_t \rightarrow R_t$.
- The actor modifies its policy when feedbacks do not correspond to predictions.

Reward Prediction Error

Should the Critic predict correctly, we should have:

$$\begin{array}{rrrrrr} P_{t-1} = & R_{t-1} = & r_t + & \gamma r_{t+1} + & \gamma^2 r_{t+2} + & \gamma^3 r_{t+3} + \ldots \\ P_t = & R_t = & & r_{t+1} + & \gamma r_{t+2} + & \gamma^2 r_{t+3} + \ldots \end{array}$$

thus, we should have:

$$P_{t-1} = r_t + \gamma P_t$$

if not, there is a reward prediction error (RPE):

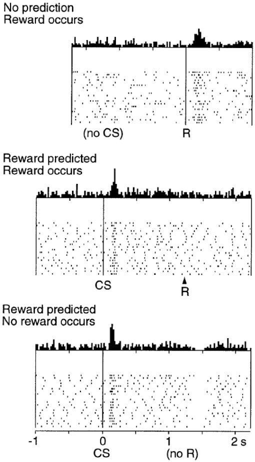
$$\delta = r_t + \gamma P_t - P_{t-1}$$

If $\delta < 0$, predictions should be decreased (C), and probability of last action selection should decrease (A).
If $\delta > 0$, predictions should be increased (C), and probability of last action selection should increased (A).

What about the brain? (Schultz et al., 1997)

Dopaminergic neurons

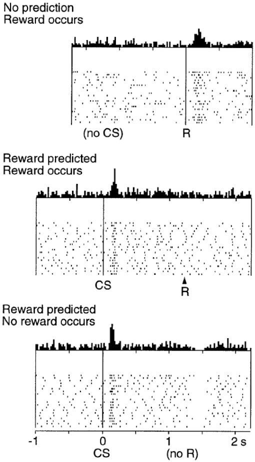
$$\delta_t = r_t + \gamma P_t - P_{t-1}$$



- R : $r_t = 0$ expected, $P_{t-1} = \gamma P_t$
  $\delta = R$
- CS : unpredictable stimulus
  $\delta = R$
- R : $r_t = R$ expected,
  $P_{t-1} = R + \gamma P_t$
  $\delta = 0$
- CS : $\delta = R$
- R : $r_t = R$ expected,
  $P_{t-1} = R + \gamma P_t$
  $\delta = -R$

What about the brain? (Schultz et al., 1997)

Dopaminergic neurons



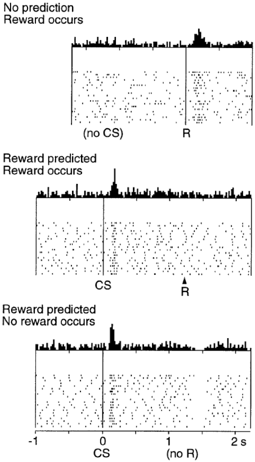
$$\delta_t = r_t + \gamma P_t - P_{t-1}$$

- R : $r_t = 0$ expected, $P_{t-1} = \gamma P_t$
  $\delta = R$
- CS : unpredictable stimulus
  $\delta = R$
- R : $r_t = R$ expected,
  $P_{t-1} = R + \gamma P_t$
  $\delta = 0$
- CS : $\delta = R$
- R : $r_t = R$ expected,
  $P_{t-1} = R + \gamma P_t$
  $\delta = -R$

What about the brain? (Schultz et al., 1997)

Dopaminergic neurons

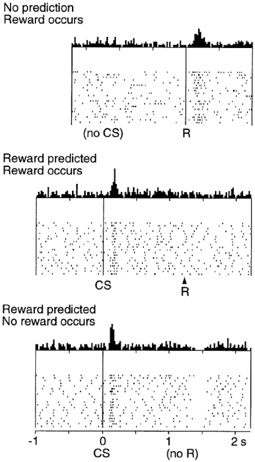
$$\delta_t = r_t + \gamma P_t - P_{t-1}$$



- R : $r_t = 0$ expected, $P_{t-1} = \gamma P_t$
  $\delta = R$
- CS : unpredictable stimulus
  $\delta = R$
- R : $r_t = R$ expected,
  $P_{t-1} = R + \gamma P_t$
  $\delta = 0$
- CS : $\delta = R$
- R : $r_t = R$ expected,
  $P_{t-1} = R + \gamma P_t$
  $\delta = -R$

What about the brain? (Schultz et al., 1997)

Dopaminergic neurons



$$\delta_t = r_t + \gamma P_t - P_{t-1}$$

- R : $r_t = 0$ expected, $P_{t-1} = \gamma P_t$
  $\delta = R$
- CS : unpredictable stimulus
  $\delta = R$
- R : $r_t = R$ expected,
  $P_{t-1} = R + \gamma P_t$
  $\delta = 0$
- CS : $\delta = R$
- R : $r_t = R$ expected,
  $P_{t-1} = R + \gamma P_t$
  $\delta = -R$

Model based-RL

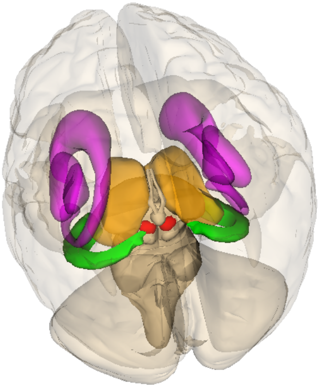
If the agent tries to build a model of the world:

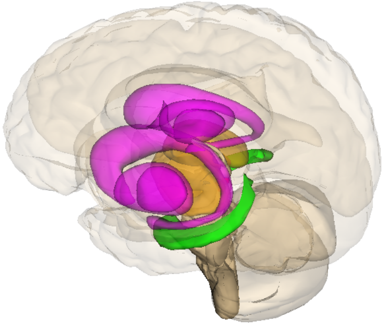
- **reward model**: which states provide rewards or punishments?
- **transition model**: in which state do you end-up after doing action a in state s?

It can be exploited to directly estimate the values of states and the optimal policy (with a process akin to planning).

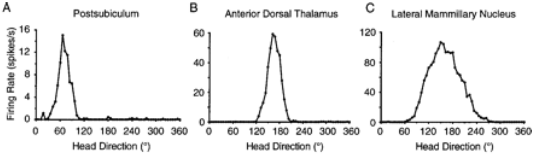
(more details to come)

Neural substrate of Navigation

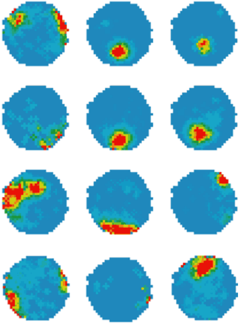


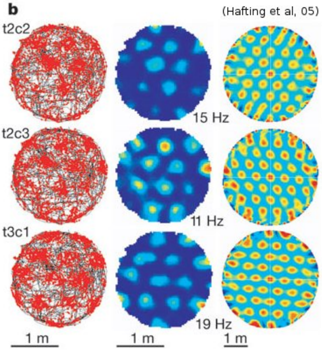


Neural substrate of Navigation



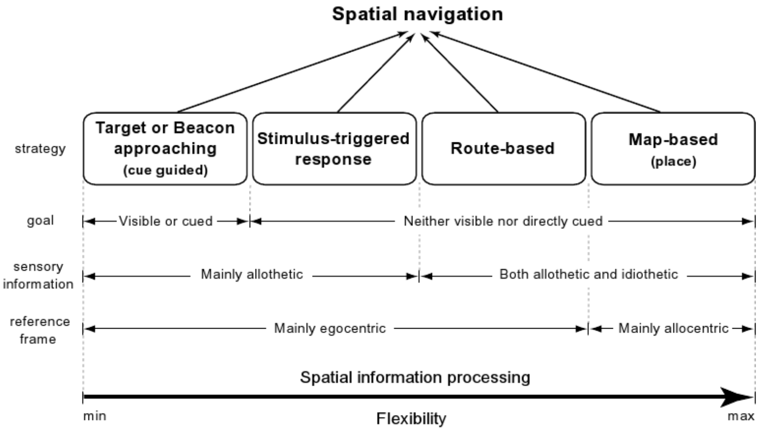
(Taube et al., Cereb Cortex, 2003, 13:1162-1172)





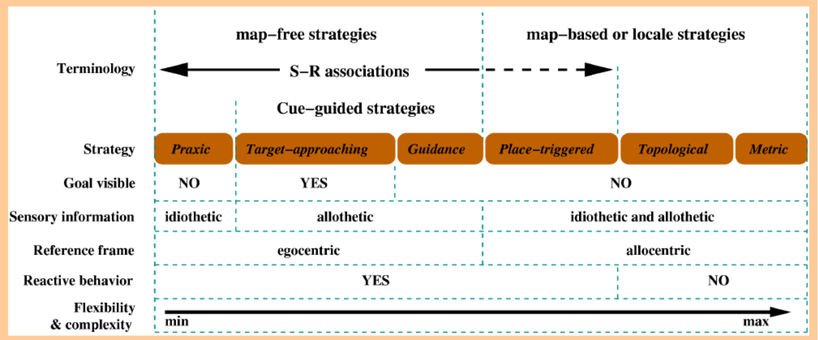
Plan

Complexity/Flexibility



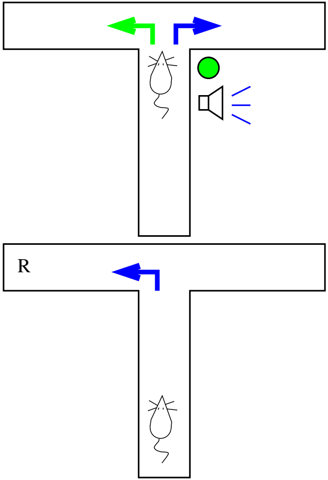
(Arleo & Rondi-Reig, 2007)

model-free/model-based $\neq$ map-based/map-free



(Khamassi, 2007)

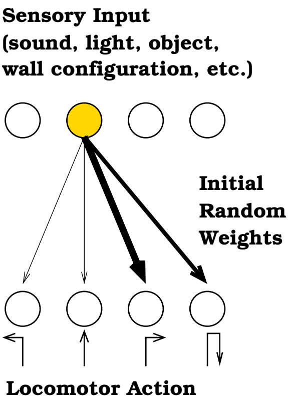
Stimulus triggered response



characteristics

- a stimulus
  $\Rightarrow$ an action,
- model-free RL.

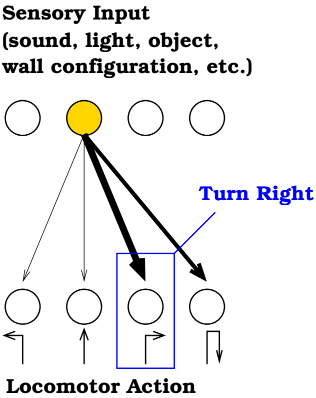
Stimulus triggered response



characteristics

- a stimulus
  $\Rightarrow$ an action,
- model-free RL.

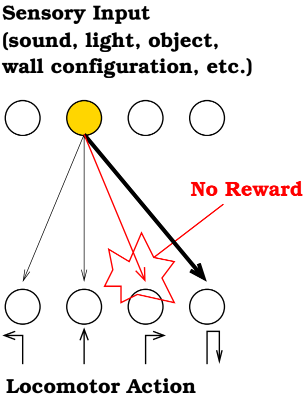
Stimulus triggered response



characteristics

- a stimulus
  $\Rightarrow$ an action,
- model-free RL.

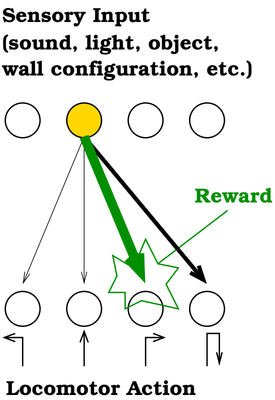
Stimulus triggered response



characteristics

- a stimulus
  $\Rightarrow$ an action,
- model-free RL.

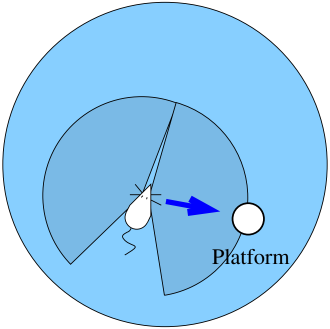
Stimulus triggered response



characteristics

- a stimulus
  $\Rightarrow$ an action,
- model-free RL.

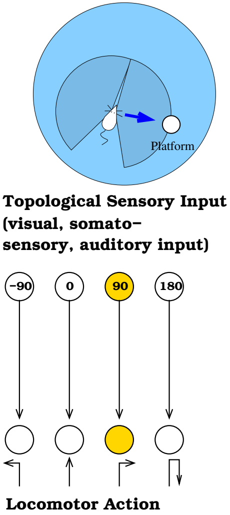
Target approach



Characteristics

- target visible (US) $\Rightarrow$ pre-wired motor response,
- calibration : supervised learning.
- neural substrate: superior colliculus (Felsen & Mainen, 2008).

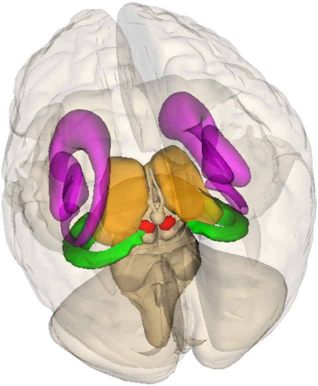
Target approach



Characteristics

- target visible (US) ⇒ pre-wired motor response,
- calibration : supervised learning.
- neural substrate: superior colliculus (Felsen & Mainen, 2008).

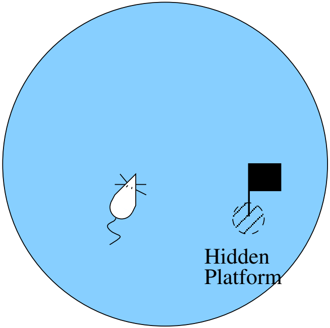
Target approach



Characteristics

- target visible (US)
  ⇒ pre-wired motor response,
- calibration :
  supervised learning.
- neural substrate:
  superior colliculus (Felsen & Mainen, 2008).

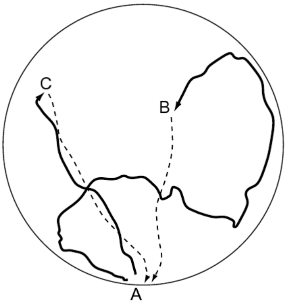
Cue approach



Characteristics

- cue visible (CS)
  ⇒ learn to select the relevant sensory information
  ⇒ no need to learn motor response,
- sensory information filtering: model free RL.

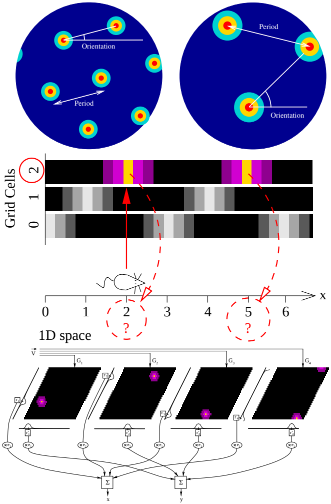
Path integration



Characteristics

- integration wrt. an origin,
- inversion: direct return path,
- no learning,
- mechanism no well known yet, involves the grid cells (Hafting et al., 2005),
- wich encode position (Fiete et al. 2008, Masson & Girard, 2009).
- **Integration of movements: accumulates errors**

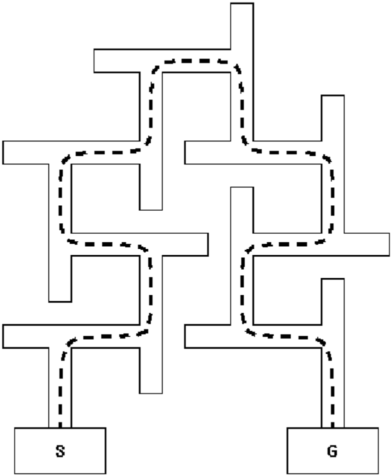
Path integration



Characteristics

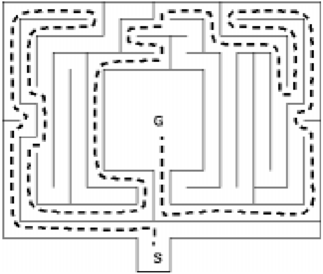
- integration wrt. an origin,
- inversion: direct return path,
- no learning,
- mechanism no well known yet, involves the grid cells (Hafting et al., 2005),
- wich encode position (Fiete et al. 2008, Masson & Girard, 2009).
- **Integration of movements: accumulates errors**

Praxic strategy

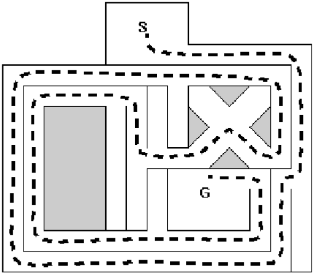


Characteristics

- (Watson, 1907; Honzic, 1936) : blind, deaf rats, without smell and whiskers learn to solve the maze *without touching walls*.



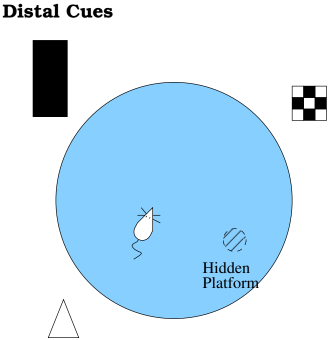
Praxic strategy



Characteristics

- (Watson, 1907; Honzic, 1936) : blind, deaf rats, without smell and whiskers learn to solve the maze *without touching walls*.
- (Carr & Watson, 1908) : they hit the wall if the corridors are shortened.
- can be learn by imitation of another strategy (Hebbian learning sufficient).

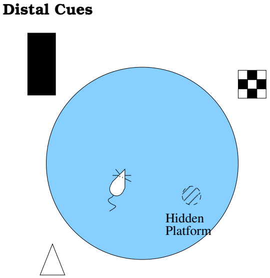
Place Recognition Triggered Response

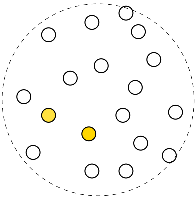


Characteristics

- Place $\Rightarrow$ action
- place representation.
- model-free RL (same algorithm, different inputs).

Place Recognition Triggered Response

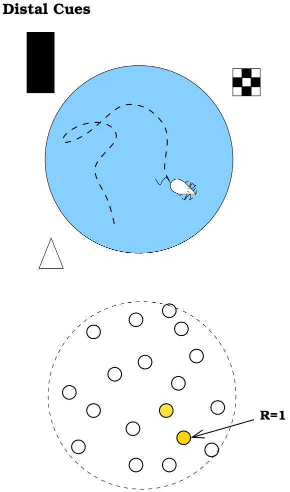




Characteristics

- Place ⇒ action
- place representation.
- model-free RL (same algorithm, different inputs).

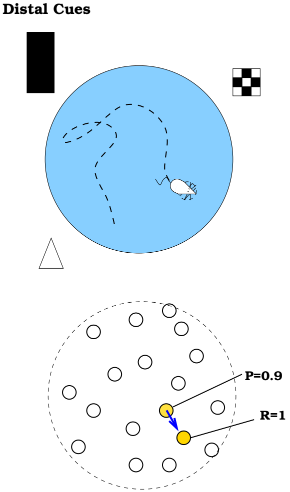
Place Recognition Triggered Response



Characteristics

- Place $\Rightarrow$ action
- model-free RL (same algorithm, different inputs).
- Slow to converge.

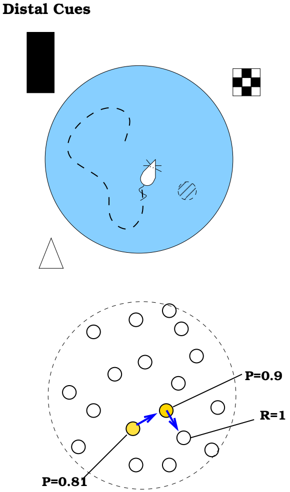
Place Recognition Triggered Response



Characteristics

- Place ⇒ action
- model-free RL (same algorithm, different inputs).
- Slow to converge.

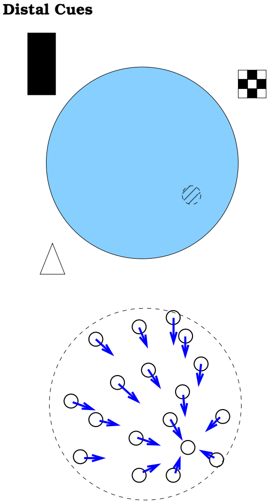
Place Recognition Triggered Response



Characteristics

- Place ⇒ action
- model-free RL (same algorithm, different inputs).
- Slow to converge.

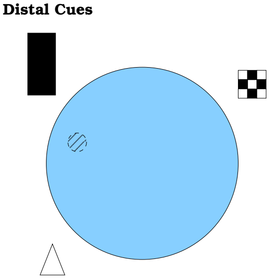
Place Recognition Triggered Response

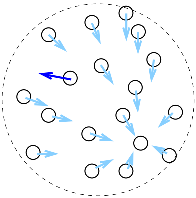


Characteristics

- Place ⇒ action
- model-free RL (same algorithm, different inputs).
- Slow to converge.

Place Recognition Triggered Response

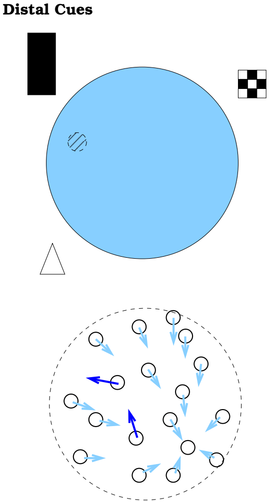




Characteristics

- Place ⇒ action
- model-free RL (same algorithm, different inputs).
- Slow to converge.
- Difficult to adapt to changes.

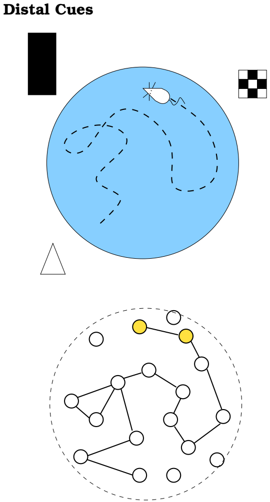
Place Recognition Triggered Response



Characteristics

- Place $\Rightarrow$ action
- model-free RL (same algorithm, different inputs).
- Slow to converge.
- Difficult to adapt to changes.

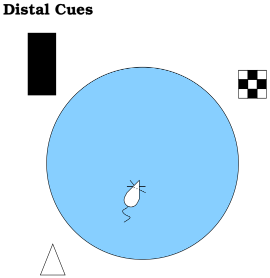
Planning

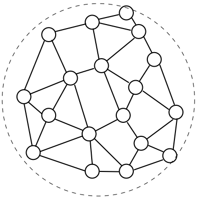


Characteristics

- Build a world-model: reward and transition functions.
- transitions can be learnt latently.

Planning

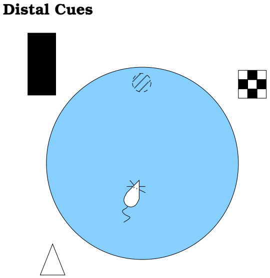


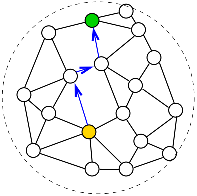


Characteristics

- Build a world-model: reward and transition functions.
- transitions can be learnt latently.

Planning

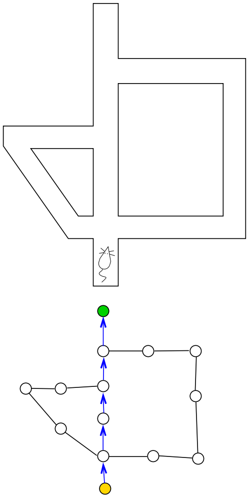




Characteristics

- Build a world-model: reward and transition functions.
- transitions can be learnt latently.
- computation-heavy planning.
- very adaptive to changes.

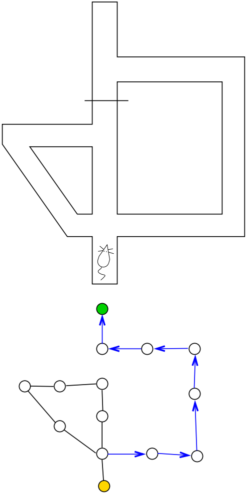
Planning



Characteristics

- Build a world-model: reward and transition functions.
- transitions can be learnt latently.
- computation-heavy planning.
- very adaptive to changes.

Planning



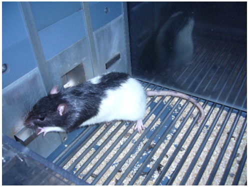
Characteristics

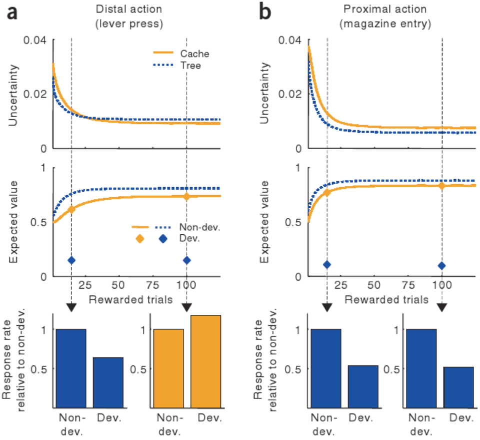
- Build a world-model: reward and transition functions.
- transitions can be learnt latently.
- computation-heavy planning.
- very adaptive to changes.

Plan

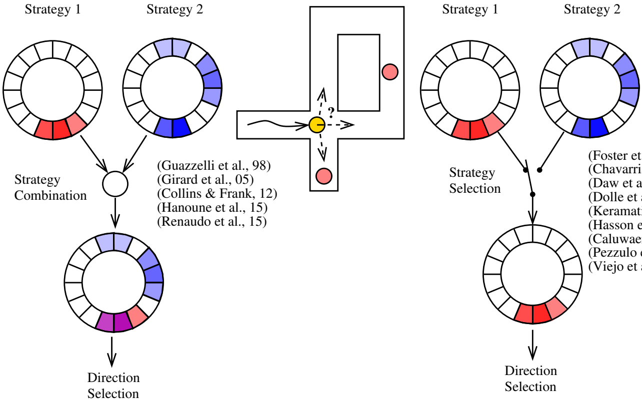
Model-based & Model-free learning algorithms

Interactions of model-based and model-free learning algorithms to explain intrumental conditioning (Daw et al., 2005, Nat. Neurosci. ; Keramati et al., 2011, PLoS Comput. Biol.).





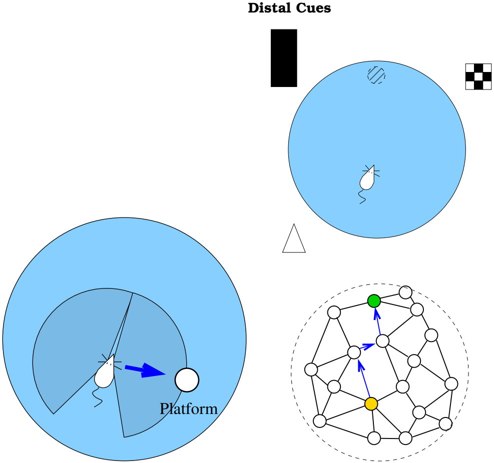
Coordination of multiple RL systems: fusion or selection?



(Guazzelli et al., 98)
(Girard et al., 05)
(Collins & Frank, 12)
(Hanoune et al., 15)
(Renaudo et al., 15)

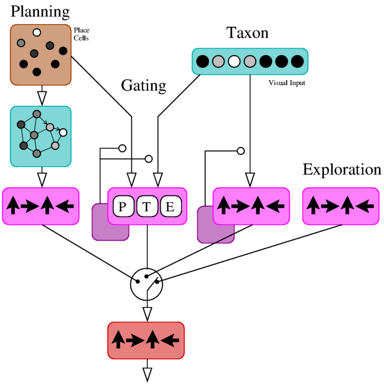
(Foster et al., 00)
(Chavarriaga et al., 05)
(Daw et al., 05)
(Dolle et al., 10)
(Keramati et al., 11)
(Hasson et al., 11)
(Caluwaerts et al., 12a,b)
(Pezzulo et al., 13)
(Viejo et al., 15)

(Dollé et al., 2010): Strategies



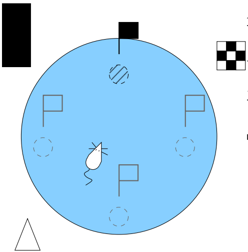
(Dollé et al., 2010, *Biological Cybernetics*, 103(4):299–317)

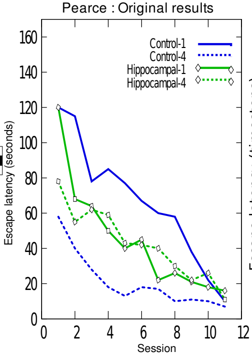
(Dollé et al., 2010): Arbitration mechanism

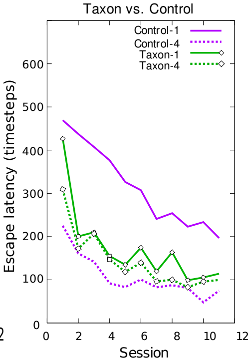


- parallel neural substrates,
- adaptive coordination (model-free RL),
- combines different learning algorithms (model-based, model-free, etc.),
- exhibits cooperation and competition,
- exploration regulation.

Reproduction of (Pearce et al., 1998)

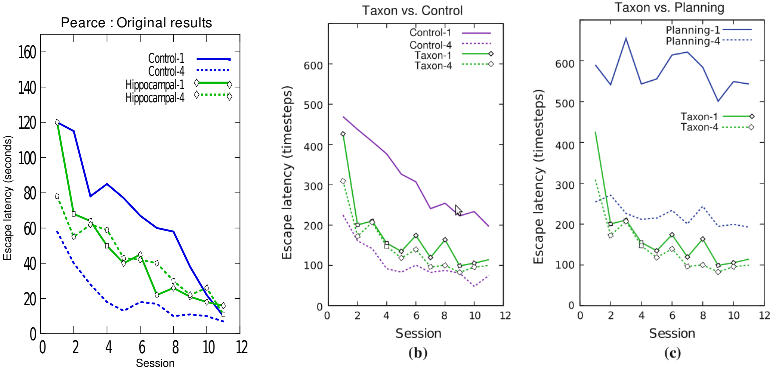






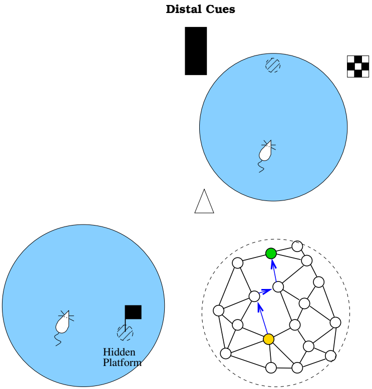
4 trials, 11 sessions. Control vs. hippocampal rats.

Reproduction of (Pearce et al., 1998)



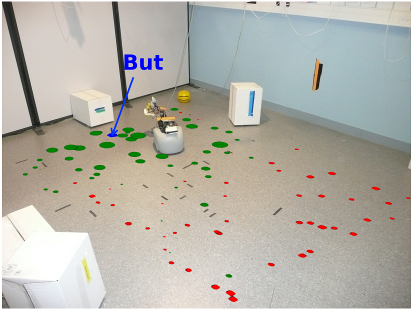
4 trials, 11 sessions. Control vs. hippocampal rats.

(Caluwaerts et al., 12a,b): Strategies



(Caluwaerts, Staffa, N'Guyen, Grand, Dollé, Favre-Felix, Girard & Khamassi (2012). Bioinspiration & Biomimetics. Vol 7(2):025009.)
(Caluwaerts, Favre-Felix, Staffa, N'Guyen, Grand, Girard & Khamassi, (2012). Living Machines 2012, LNAI 7375/2012, p. 62-73.)

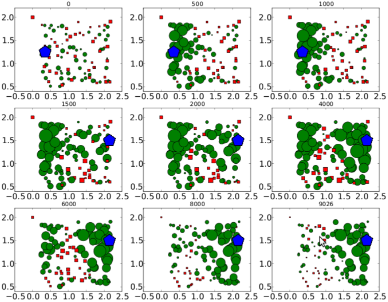
(Caluwaerts et al., 12a,b): Results

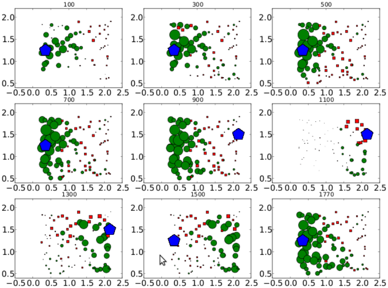


Results

Appropriate strategy selection wrt. efficiency.
Context detection algorithm for an enhanced adaptation to task changes.

(Caluwaerts et al., 12a,b): Results

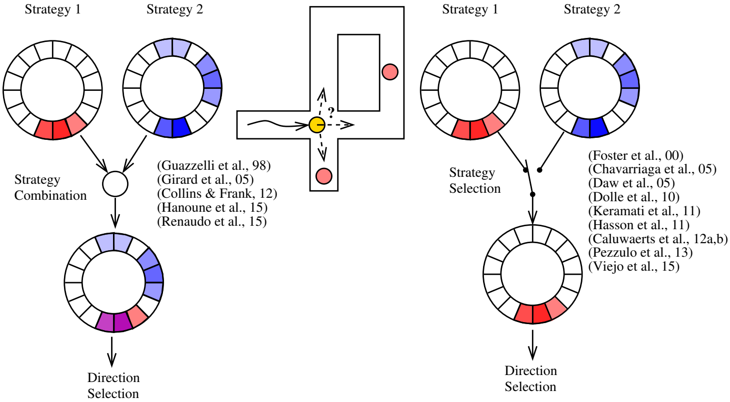




Results

Appropriate strategy selection wrt. efficiency.
Context detection algorithm for an enhanced adaptation to task changes.

Coordination of multiple RL systems: fusion or selection?



Coordination criteria in current models

Coordination: predetermined (e.g. Guazzelli, Girard) or adaptive (e.g. Foster, Chavarriaga, Dollé).
Criteria :

- Reward prediction,
- Reward prediction error,
- Estimated uncertainty.

BUT few strategies involved in general (2-3)
To be explored:

- Changes in average reward rates,
- Entropy of value distributions, and evolution,
- Computational cost, etc.

Plan

Wrap-up

Take-home messages

- Multiple RL algorithm families have been developed in AI.
- They appear to be good models of animal behavior (& links with neural substrate).
- The exact operation and the neural substrate of multiple decision systems coordination are still unknown.
- All RL algorithms are useful to explain navigation behaviors.
- BUT Navigation tells us that RL is not the only way to make a decision.

Bibliography

- Arleo & Gerstner (2000). Spatial cognition and neuro-mimetic navigation: a model of hippocampal place cell activity. *Biological Cybernetics*, 83(3):287–300.
- Arleo & Rondi-Reig (2007). Multimodal sensory integration and concurrent navigation strategies for spatial cognition in real and artificial organisms. *J Integr Neurosci*. 6(3):327–366.
- Caluwaerts, Staffa, N'Guyen, Grand, Dollé, Favre-Felix, Girard & Khamassi (2012a). A biologically inspired meta-control navigation system for the Psikharpax rat robot. *Bioinspiration & Biomimetics*. Vol 7(2):025009.
- Caluwaerts, Favre-Felix, Staffa, N'Guyen, Grand, Girard & Khamassi (2012b). Neuro-inspired navigation strategies shifting for robots: Integration of a multiple landmark taxon strategy. *Living Machines 2012*, Prescott, T.J. et al. (Eds.). LNAI 7375/2012, Pages 62-73.
- Chavarriaga, Strösslin, Sheynikhovich & Gerstner (2005). A Computational Model of Parallel Navigation Systems in Rodents. *Neuroinformatics*. 3(3):223–242.
- Daw, Niv & Dayan (2005). Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioral control. *Nature Neurosci*. 8(12):1704–1711.
- Dollé, Khamassi, Girard, Guillot & Chavarriaga (2008). Analyzing interactions between navigation strategies using a computational model of action selection. In *Spatial Cognition VI*. LNCS:71–86, Springer.
- Dollé, Sheynikhovich, Girard, Chavarriaga, Guillot (2010a). Path planning versus cue responding: a bioinspired model of switching between navigation strategies. *Biological Cybernetics*. 103(4):299–317.
- Dollé, Sheynikhovich, Girard, Ujfalussy, Chavarriaga, & Guillot (2010b). Analyzing interactions between cue-guided and place-based navigation with a computational model of action selection: Influence of sensory cues and training. *From animals to animats 11*, Springer. LNAI 6226:335–346.
- Fiete, Burak, & Brookings (2008). What grid cells convey about rat location. *J Neurosci*, 28(27):6858.
- Foster, Morris & Dayan (2000). Models of Hippocampally Dependent Navigation using the Temporal Difference Learning Rule. *Hippocampus*. 10:1–16.
- Gaussier, Revel, Banquet, Babeau (2002). From view cells and place cells to cognitive map learning: processing stages of the hippocampal system. *Biological Cybernetics* 86(1):15–28.
- Girard, Filliat, Meyer, Berthoz & Guillot (2005). Integration of navigation and action selection functionalities in a computational model of cortico-basal ganglia-thalamo-cortical loops. *Adaptive Behavior*, 13(2):115–130.

Bibliography

- Guazzelli, Corbacho, Bota & Arbib (1998). Affordances, Motivation, and the World Graph Theory. *Adaptive Behavior*. 6(3/4):435–471.
- Jaeger & Hass (2004). Harnessing Nonlinearity: Predicting Chaotic Systems and Saving Energy in Wireless Communication. *Science* 304(5667):78–80.
- Khamassi, Lachèze, Girard, Berthoz, & Guillot (2005). Actor-critic models of reinforcement learning in the basal ganglia: From natural to artificial rats. *Adaptive Behavior*. 13(2):131-148.
- Khamassi, Martinet, Guillot (2006) Self-organising maps with mixture of exprets: Application to an Actor-Critic Model of reinforcement Learning in the basal Ganglia, In *SAB 2006*, Springer Verlag.
- Khamassi (2007). Rôles complémentaires du cortex préfrontal et du striatum dans l'apprentissage et le changement de stratégies de navigation fondées sur la récompense chez le rat. Thèse de doctorat UPMC.
- Martinet, Fouque, Passot, Meyer & Arleo (2008). Modelling the cortical columnar organisation for topological state-space representation, and action planning, In *SAB 2008*, 5040:137–147, Springer-Verlag.
- Masson and Girard (2009). Decoding the Grid Cells for Metric Navigation Using the Residue Numeral System. *ICCN2009*. Hangzhou, P.R. China.
- Sussillo & Abbott (2009). Generating Coherent Patterns of Activity from Chaotic Neural Networks. *Neuron* 63:544–557.