Reinforcement Learning & the Brain

Mehdi Khamassi
CogMaster CA10, Course #4, 15 Mar 2019
**RESEARCH INTERESTS**

- **Decision-making**: choice at each moment of the most appropriate behavior for an agent’s survival, to solve a task.

- **Reinforcement Learning** (by trial/error): adaptation of this choice to maximize a particular reward function (usually the sum of cumulative reward over an infinite horizon in a Markov Decision Process):  
  \[ f(t) = \sum_{t=0}^{\infty} \gamma^t r_t \ (\gamma < 1) \]

- **Complex problems**: noise, partial representation of states, non stationarity of the environment.

- **Modular/hierarchical structure** of different learning levels, enabling a better flexibility and autonomy of decision in animals and robots. Research strategy: Bio/Neuro inspiration
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.
Learning rate (\(=0.9\))
Discount factor (\(=0.9\))

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

\[ V(s_t) = V(s_t) + \alpha \cdot \delta_{t+1} \]
REINFORCEMENT LEARNING in a Markov Decision Process

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

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

discount factor (=0.9)

learning rate (=0.9)
REINFORCEMENT LEARNING
in a Markov Decision Process

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

\[ 0.9 = 0 + 0.9 \cdot 1 \]

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

\[ 1 = 1 + 0 - 0 \]

discount factor (=0.9)

learning rate (=0.9)
REINFORCEMENT LEARNING in a Markov Decision Process

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

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

Color indicates value.

Learning rate (\(\alpha\)) = 0.9
Discount factor (\(\gamma\)) = 0.9
REINFORCEMENT LEARNING in a Markov Decision Process

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

\[
0 = 0 + 0.9 \cdot 0
\]

discount factor (=0.9)

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

learning rate (=0.9)
REINFORCEMENT LEARNING
in a Markov Decision Process

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

\[ 0.81 = 0 + 0.9 \times 0.9 - 0 \]

\[ 0.72 = 0 + 0.9 \times 0.81 \]

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

Discount factor (=0.9)

Learning rate (=0.9)
REINFORCEMENT LEARNING in a Markov Decision Process

\[ 0.81 = 0 + 0.9 \times 0.9 - 0 \]

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

\[ 0.72 = 0 + 0.9 \times 0.81 \]

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

Color indicates value

discount factor (=0.9)

learning rate (=0.9)
REINFORCEMENT LEARNING in a Markov Decision Process

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

$0.1 = 1 + 0 - 0.9$

$0.99 = 0.9 + 0.9 \times 0.1$

discount factor (=0.9)

learning rate (=0.9)
REINFORCEMENT LEARNING
in a Markov Decision Process

\[ 0.1 = 1 + 0 - 0.9 \]

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

\[ 0.99 = 0.9 + 0.9 \times 0.1 \]

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

discount factor (\(=0.9\))

learning rate (\(=0.9\))

Color indicates value
**REINFORCEMENT LEARNING**

in a Markov Decision Process

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

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

- After \(N\) simulations
- Very long!

**Discount factor** (=0.9)

**Learning rate** (=0.1)

usually small for stability
REINFORCEMENT LEARNING
in a Markov Decision Process

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

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

After N simulations

Very long!

discount factor (=0.9)

learning rate (=0.1)
REINFORCEMENT LEARNING in a Markov Decision Process

After 
N simulations
Very long!

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

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

---

discount factor (\(=0.9\))

learning rate (\(=0.1\))
REINFORCEMENT LEARNING in a Markov Decision Process

May converge to a sub-optimal solution!

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

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

- discount factor (=0.9)
- learning rate (=0.1)
REINFORCEMENT LEARNING in a Markov Decision Process

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

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

Exploration-Exploitation trade-off

discount factor (=0.9)

learning rate (=0.1)
REINFORCEMENT LEARNING in a Markov Decision Process

Finds best solution after infinite time!

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

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

Discount factor (=0.9)

Learning rate (=0.1)
Links with biology

*Activity of dopaminergic neurons*
TD-learning explains classical conditioning (predictive learning)

**Before Conditioning**
- UCS (food in mouth) → UCR (salivation)
- Neutral stimulus (tone) → No salivation

**During Conditioning**
- Neutral stimulus (tone) + UCS (food in mouth) → UCR (salivation)

**After Conditioning**
- The neutral stimulus alone now produces a conditioned response (CR), thereby becoming a conditioned stimulus (CS).

Taken from Bernard Balleine’s lecture at Okinawa Computational Neuroscience Course (2005).
REINFORCEMENT LEARNING

- Analogy with dopaminergic neurons’ activity

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

Schultz et al. (1997); Bayer & Glimcher (2005); Morris et al. (2006); Roesch et al. (2007); Matsumoto & Hikosaka (2009).
REINFORCEMENT LEARNING

- Analogy with dopaminergic neurons’ activity

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

Schultz et al. (1993); Houk et al. (1995); Schultz et al. (1997).
Analogy with dopaminergic neurons’ activity

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

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

- Analogy with dopaminergic neurons’ activity

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

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

Brain responses

RL model

Prediction error

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

$S$: state space

$A$: action space

Policy function $\pi: S \rightarrow A$

What we have learned so far:

Value function $V: S \rightarrow \mathbb{R}$
How can the agent learn a policy?

*How to learn to perform the right actions*

A solution: parallelly update a policy and a value function

\[
\pi(a_t|s_t) = \pi(a_t|s_t) + \alpha \cdot \delta_{t+1}
\]

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

\[ Q : (S,A) \rightarrow \mathbb{R} \]

\[ Q \text{-table:} \]

<table>
<thead>
<tr>
<th>state / action</th>
<th>a1 : North</th>
<th>a2 : South</th>
<th>a3 : East</th>
<th>a4 : West</th>
</tr>
</thead>
<tbody>
<tr>
<td>s1</td>
<td>0.92</td>
<td>0.10</td>
<td>0.35</td>
<td>0.05</td>
</tr>
<tr>
<td>s2</td>
<td>0.25</td>
<td>0.52</td>
<td>0.43</td>
<td>0.37</td>
</tr>
<tr>
<td>s3</td>
<td>0.78</td>
<td>0.9</td>
<td>1.0</td>
<td>0.81</td>
</tr>
<tr>
<td>s4</td>
<td>0.0</td>
<td>1.0</td>
<td>0.9</td>
<td>0.9</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>
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)

\[ Q : (S,A) \rightarrow \mathbb{R} \]

**Q-table**:

<table>
<thead>
<tr>
<th>state / action</th>
<th>a1 : North</th>
<th>a2 : South</th>
<th>a3 : East</th>
<th>a4 : West</th>
</tr>
</thead>
<tbody>
<tr>
<td>s1</td>
<td>0.92</td>
<td>0.10</td>
<td>0.35</td>
<td>0.05</td>
</tr>
<tr>
<td>s2</td>
<td>0.25</td>
<td>0.52</td>
<td>0.43</td>
<td>0.37</td>
</tr>
<tr>
<td>s3</td>
<td>0.78</td>
<td>0.9</td>
<td>1.0</td>
<td>0.81</td>
</tr>
<tr>
<td>s4</td>
<td>0.0</td>
<td>1.0</td>
<td>0.9</td>
<td>0.9</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

![Diagram of a grid world with Q-values](image)
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)

\[ Q : (S,A) \rightarrow \mathbb{R} \]

\[
P(a) = \frac{\exp(\beta \cdot Q(s,a))}{\sum_b \exp(\beta \cdot Q(s,b))}
\]

The \( \beta \) parameter regulates the exploration – exploitation trade-off.
Reward Prediction Error in all Temporal-Difference (TD) methods

- **V-LEARNING (e.g. ACTOR-CRITIC)**

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

State-dependent Reward Prediction Error

(independent from the action)
Reward Prediction Error in all Temporal-Difference (TD) methods

- **V-LEARNING (e.g. ACTOR-CRITIC)**

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

*State-dependent Reward Prediction Error*

(independent from the action)

\[
P(a_t|s_t) \leftarrow P(a_t|s_t) + \alpha \delta_{t+1}
\]

Also used to update the ACTOR
Reward Prediction Error in all Temporal-Difference (TD) methods

- **V-LEARNING** (e.g. 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
Reward Prediction Error in all Temporal-Difference (TD) methods

- **V-LEARNING** *(e.g. 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**

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

Reward Prediction Error dependent on the best action
Does dopamine still encode a reward prediction error when the animal has to perform a choice?

Corollary: Which TD method best describes the information carried by dopamine phasic responses?
Two contradictory findings by Electrophysiological experiments

- Neural recordings done in monkeys (Morris et al. 2006) and rats (Roesch et al. 2007) performed to investigate which TD algorithm/model (Actor-Critic vs Q-learning vs SARSA) best describe the information encoded in dopamine signals.
  - Roesch et al (2007) found that it reflects the best possible future action (Q-LEARNING).
Typical probabilistic decision-making task

Typical probabilistic decision-making task

Typical probabilistic decision-making task

\[ P(\text{left choice}) = 0.2 \]
\[ V(\square) = 0.85 \]
\[ Q(\square, \text{L}) = 0.25 \]
\[ P(\text{right choice}) = 0.8 \]
\[ V(\circ) = 0.85 \]
\[ Q(\circ, \text{R}) = 1 \]

Prediction error (dopamine)

V learning
Q learning
SARSA

Contradictory finding: Dopamine neurons encode the better option
Another report in rats concludes in favor of Q-learning.
Daw (2007), commentary about the results presented in Roesch et al. (2007).
Dopamine neurons encode the better option in rats (Roesch 2007)

Same amplitude no matter which action (Q-learning)
Dopamine neurons encode the better option in rats (Roesch 2007)

Same amplitude no matter which action (Q-learning)
Dopamine neurons encode the better option in rats (Roesch 2007)

But: Problems!

Value function rather than reward prediction error?
Model-based analysis
Work by Jean Bellot (PhD student)

Model-based analysis
Work by Jean Bellot (PhD student)

Model-based analysis
Work by Jean Bellot (PhD student)

Model

Neural activity

Signal averaged over all post-learning trials (as in original exp.)

Signal averaged over the 9 first post-learning trials

Model-based analysis
Work by Jean Bellot (PhD student)

Model comparison

Bellot et al. (submitted)
SUMMARY OF MODEL FITTING ON DOPAMINE PHASIC ACTIVITY:

Parameters fitted on the rat’s behavior ($\alpha=0.3$) differ from those that best describe dopaminergic activity ($\alpha=0.1$).

Dopamine activity better fitted by a mixture of reward prediction error and value! (see Cohen et al. 2012, Nature)
Model-based analysis
Work by Jean Bellot (2011)

- Parameters fitted on the rat’s behavior differ from those that best describe dopaminergic activity
→ Idea that behavior is not completely linked to learning dynamics reflected in dopamine activity.
→ Idea that behavior might be the result of parallel learning systems (Daw et al., 2005)

Diagram:
- Q-learning
- Other learning system (?)
- competition / cooperation
- behavior
REINFORCEMENT LEARNING

After N simulations

Very long!

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

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

discount factor (=0.9)  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).

\[ S : \text{state space} \]
\[ A : \text{action space} \]

Transition function \( T : S \times A \rightarrow S \)

Reward function \( R : S \times A \rightarrow R \)
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)
s : state of the agent (●)

a : action of the agent (go east)

stored transition function $T$:

$\text{proba}(\rightarrow) = 0.9$

$\text{proba}(\rightarrow) = 0.1$

$\text{proba}(\leftarrow) = 0$
Model-based Reinforcement Learning

s : state of the agent (●)

a : action of the agent (go east)

stored transition function $T$:

$\text{proba}(\rightarrow) = 0.9$

$\text{proba}(\leftarrow) = 0.1$

$\text{proba}(\uparrow) = 0$

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

$0.6 \quad 0 \quad 0.9 \times 0.7 + 0.1 \times 0.9 + 0 \times 0.3 + \ldots$
No reward prediction error!

Only:
- Estimated Q-values
- Transition function
- Reward function

This process is called Value Iteration or Dynamic prog.

\[
Q(s, a) \leftarrow R(s, a) + \gamma \sum_{s'} T(s' | s, a) \max_{a'} Q(s', a')
\]
Model-based Reinforcement Learning

Links with neurobiological data

Activity of hippocampal place neurons
Hippocampal place cells

- Reactivation of hippocampal place cells during sleep (Wilson & McNaughton, 1994, Science)
- Forward replay of hippocampal place cells during sleep (sequence is compressed 7 times) (Euston et al., 2007, Science)
a) Labyrinthe à 8 bras dont trois récompensés

b) Sommeil/Repos

<table>
<thead>
<tr>
<th>Test</th>
<th>Contrôle</th>
</tr>
</thead>
<tbody>
<tr>
<td>exemple de &quot;ripple&quot;</td>
<td>exemple de &quot;ripple&quot;</td>
</tr>
<tr>
<td>&quot;ripple&quot; interrompu</td>
<td>&quot;ripple&quot; intact</td>
</tr>
</tbody>
</table>

C) Index de Performance

<table>
<thead>
<tr>
<th>Groupe contrôle stimulé</th>
<th>Groupe contrôle non stimulé</th>
<th>Groupe test stimulé</th>
</tr>
</thead>
<tbody>
<tr>
<td>Jours 1</td>
<td>Jours 2</td>
<td>Jours 3</td>
</tr>
<tr>
<td>0.4</td>
<td>0.5</td>
<td>0.6</td>
</tr>
</tbody>
</table>
A. Spatial right (or left) vs. Cue guided light (or dark)

B. Sleep Pre (~30 min) | Task (~45 min) | Sleep Post (~30 min)

Time
Reactivations are stronger for learning sessions


"Decision point": place of high coherence between PFC and HIP


Distinct PFC populations for task change detection and strategy selection

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)
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**


Cully et al. 2015 Nature
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.

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

(Daw Niv Dayan 2005, Nat Neurosci)
Habitual vs goal-directed: sensitive to changes in outcome

Yin et al. 2004; Balleine 2005; Yin & Knowlton 2006
Progressive shift from model-based navigation to model-free navigation

Model-free navigation

Model-based navigation

Benoît Girard 2010 UPMC lecture
Work by Laurent Dollé:

Dollé Chavarriaga Guillot Khamassi, 2018, PLoS Computational Biology
Dual learning in a rat-like robot (European project ICEA)

Model-based (MB) strategy only

MB strategy + MF strategy

Caluwaerts et al. (2012) Biomimetics & Bioinspiration
Behavioral results

Sign-trackers

Goal-trackers

Fast Scan Cyclic Voltammetry (FSCV) in the ventral striatum.

Systemic injection of flupentixol prior to each session.

Modelling the task as a Markov Decision Process

\[
\begin{aligned}
(1 - \omega) & \quad \text{MB} \\
\mathcal{T} & \xrightarrow{\mathcal{R}} A \\
+ \quad \omega & \quad \text{FMF} \\
\end{aligned}
\]

\[
\begin{aligned}
\mathcal{P}_{goL}(s_1) & = (1 - \omega) \left( Q_{goL}(s_1) - \max_{a'} Q_{a'}(s_1) \right) + \omega \quad \mathcal{V}(L) \\
\mathcal{P}_{goM}(s_1) & = (1 - \omega) \left( Q_{goM}(s_1) - \max_{a'} Q_{a'}(s_1) \right) + \omega \quad \mathcal{V}(M)
\end{aligned}
\]

with $\omega = 0.499$ (STs), $\omega = 0.048$ (GTs), $\omega = 0.276$ (IGs)


\[
\begin{aligned}
\mathcal{P}_{goL}(s_1) &= (1 - \omega) \left( Q_{goL}(s_1) - \max_{a'} Q_{a'}(s_1) \right) + \omega \mathcal{V}(L) \\
\mathcal{P}_{goM}(s_1) &= (1 - \omega) \left( Q_{goM}(s_1) - \max_{a'} Q_{a'}(s_1) \right) + \omega \mathcal{V}(M)
\end{aligned}
\]

with $\omega = 0.499$ (STs), $\omega = 0.048$ (GTs), $\omega = 0.276$ (IGs)
Behavioral results

Physiological results

Physiological results

Pharmacological results

Summary of the simulation results

Summary of the simulation results


Symmetrical version (impetus for L&M) of Dayan et al. 2006
Asymmetrical version (impetus for L only) of Dayan et al. 2006
**Experimental predictions**


1. **GTs more sensitive to outcome devaluation than STs.**
2. DA dip at each magazine visit during ITI.
3. DA patterns in the intermediate group.
4. **Shortening the ITI should change DA pattern in GTs.**
5. **Reducing the ITI duration should increase the tendency to goal-track in the overall population.**
Experimental predictions

1. GTs more sensitive to outcome devaluation than STs.
2. DA dip at each magazine visit during ITI.
3. DA patterns in the intermediate group.
4. Shortening the ITI should change DA pattern in GTs.
5. Reducing the ITI duration should increase the tendency to goal-track in the overall population.

New NSF-ANR (USA-France) Computational Neuroscience project (2016-2019) to test our model's predictions during experiments performed at NIH / Univ. Maryland.

Predictions 1, 4 & 5 verified!
Sign-tracking responses are not sensitive to outcome devaluation

**Preferred Response**

- Unpaired
- Paired

**Normalized Preferred Response**

- Unpaired
- Paired

*Nasser et al 2015 Frontiers in Behavioral Neuroscience*
The model assumes down-revision of unrewarded magazine value during ITI. Thus reducing ITI should lead to more GT, and increasing ITI should lead to more ST.
• Sharper version of Prediction 2: removing the magazine during ITI > reducing ITI duration.

Restoring a DA profile close to a RPE in GT.

Lengthening ITI -> more ST
Shortening ITI -> more GT

PCA index = average of response bias, probability difference, and latency difference.
• Response bias = (Lever Presses − Food Cup Entries) / (Lever Presses + Food Cup Entries)
• Probability score = P_{lever} − P_{food-cup}
• Latency index = food-cup entry latency − lever press latency
Long version (Fig. 1g-i)

(g) Average beam break at food cup (solid) and lever press (dashed) rate for 120-s (red) and 60-s (blue) ITI sessions.
(h) Same thing (lever press) with a different scale
(i) Green = food cup entries for 120s-ITI – 60s-ITI; Orange = lever pressing for 120s-ITI – 60s-ITI
Long version (Fig. 3a-d)

120s ITI group
Food Cup and Lever Press

60 s ITI group
Food Cup and Lever Press

Lever Press
Schoenbaum

• Lengthening ITI -> remaining DA peak at US both in GT & ST.

• Shortening ITI -> decrease of DA peak at US even in GT.
The model predicted that more down-revision of magazine value during ITI should result in higher surprise (RPE) at the US.
Model limitations

• The model has a fixed weight $\omega$ between MB and MF.

• The model only considers the initial behavioral response of animals at a given trial (first 4 sec).

• The model does not incorporate mechanisms for interaction between MB and MF.

• The model does not finely account for behavior during ITI.

• The model does not include information-seeking active exploration.

Lesaint et al. (2014) PLoS Computational Biology
Model generalization

- The same model could be successfully applied to account for pigeons’ light pecking behavior in a negative automaintenance procedure.

- Tuning only the $\omega$ parameter and keeping all other parameters the same, we could reproduce pigeons’ individual differences (Williams & Williams 1969; Sanabria et al 2006).

- This models pigeons as sign- and goal-trackers and predicts that pigeons do have some sort of a MB RL system.
SUMMARY

- Individual differences in autoshaping (sign-tracking and goal-tracking).

- Phasic dopamine may contribute to model-free (MF) RL mechanisms (sign-tracking). While model-based (MB) RL may be independent from phasic dopamine (goal-tracking).

- Tonic dopamine may contribute in regulating the exploration-exploitation trade-off. This could reduce/block the expression of both MF and MB systems.

- In the Pavlovian autoshaping paradigm, unrewarded access to Magazine during ITI could result in downregulation of its value.

- Hence ST/GT behaviors and associated DA patterns may also partly be influenced by environment (here ITI length)!
Combination of RL with working-memory (in humans)

Yet another example of application of multiple learning systems to experimental data in an instrumental learning task.

This time in humans.
How much of reinforcement learning is working memory, not reinforcement learning? A behavioral, computational, and neurogenetic analysis

Anne G. E. Collins and Michael J. Frank
Department of Cognitive, Linguistic and Psychological Sciences, Brown Institute for Brain Science, Brown University, Providence, RI, USA
$w_{n_5}(t+1, s) = \frac{p_{WMC}(r_t | s_t, a_t)w_{n_5}(t, s)}{p_{WMC}(r_t | s_t, a_t)w_{n_5}(t, s) + p_{RL}(r_t | s_t, a_t)(1 - w_{n_5})}$
Model comparison

Additional results:

- COMT gene related to WM capacity $N$.
- GPR6 gene related to RL learning rate.
- Schizophrenic deficits related to WM parameters.
Brovelli et al. (2011) NeuroImage
Model-free analysis of behavioral data:

A. Probability correct responses vs. Trial

- Stim 1
- Stim 2
- Stim 3

B. Reaction time (s) vs. Representative step

Brovelli et al. (2011) NeuroImage
Viejo et al. (2015) Frontiers in Behavioral Neuroscience
Viejo et al. (2015) Frontiers in Behavioral Neuroscience
Keramati et al. (2011) PLoS Computational Biology
\[
p(Deciding|t_{0 \rightarrow i}, H^{BW\text{M}}, H^{QL}) = \frac{1}{1 + \lambda_1(n - i) \exp^{-\lambda_2(2H^{max} - H^{BW\text{M}}_0 - H^{QL})}}
\]
A 1.0

Fit to RT

Fit to choice

B

Fit to RT

Fit to choice

- Bayesian Working Memory
- Entropy-based coordination
- VPI-based selection
- Weight-based mixture
- Q-Learning

Viejo et al. (2015) Frontiers in Behavioral Neuroscience
<table>
<thead>
<tr>
<th>Subject</th>
<th>-Bloc 1</th>
<th>-Bloc 2</th>
<th>-Bloc 3</th>
<th>-Bloc 4</th>
<th>All blocs</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>E-Coord</td>
<td>W-Mix</td>
<td>E-Coord</td>
<td>E-Coord</td>
<td>W-Mix</td>
</tr>
<tr>
<td>2</td>
<td>VPI-select</td>
<td>E-Coord</td>
<td>E-Coord</td>
<td>E-Coord</td>
<td>E-Coord</td>
</tr>
<tr>
<td>3</td>
<td>E-Coord</td>
<td>E-Coord</td>
<td>E-Coord</td>
<td>E-Coord</td>
<td>E-Coord</td>
</tr>
<tr>
<td>4</td>
<td>E-Coord</td>
<td>E-Coord</td>
<td>E-Coord</td>
<td>E-Coord</td>
<td>E-Coord</td>
</tr>
<tr>
<td>5</td>
<td>W-Mix</td>
<td>E-Coord</td>
<td>E-Coord</td>
<td>E-Coord</td>
<td>E-Coord</td>
</tr>
<tr>
<td>6</td>
<td>E-Coord</td>
<td>E-Coord</td>
<td>E-Coord</td>
<td>E-Coord</td>
<td>E-Coord</td>
</tr>
<tr>
<td>7</td>
<td>E-Coord</td>
<td>VPI-select</td>
<td>VPI-select</td>
<td>W-Mix</td>
<td>W-Mix</td>
</tr>
<tr>
<td>8</td>
<td>W-Mix</td>
<td>VPI-select</td>
<td>VPI-select</td>
<td>E-Coord</td>
<td>VPI-select</td>
</tr>
<tr>
<td>9</td>
<td>VPI-select</td>
<td>VPI-select</td>
<td>VPI-select</td>
<td>VPI-select</td>
<td>W-Mix</td>
</tr>
<tr>
<td>10</td>
<td>VPI-select</td>
<td>VPI-select</td>
<td>VPI-select</td>
<td>VPI-select</td>
<td>VPI-select</td>
</tr>
<tr>
<td>11</td>
<td>E-Coord</td>
<td>E-Coord</td>
<td>E-Coord</td>
<td>E-Coord</td>
<td>E-Coord</td>
</tr>
<tr>
<td>12</td>
<td>E-Coord</td>
<td>W-Mix</td>
<td>E-Coord</td>
<td>E-Coord</td>
<td>E-Coord</td>
</tr>
<tr>
<td>13</td>
<td>E-Coord</td>
<td>E-Coord</td>
<td>E-Coord</td>
<td>E-Coord</td>
<td>E-Coord</td>
</tr>
<tr>
<td>14</td>
<td>E-Coord</td>
<td>E-Coord</td>
<td>E-Coord</td>
<td>E-Coord</td>
<td>E-Coord</td>
</tr>
</tbody>
</table>

Viejo et al. (2015) Frontiers in Behavioral Neuroscience
TDRL Model
Dopamine
Model-based RL
Meta-Learning

Viejo et al. (2015) Frontiers in Behavioral Neuroscience
A. Entropy-based coordination

B. Weight-based mixture

C. VPI-based selection

D. $H(t, s_t)$

E. Trade-off

Viejo et al. (2015) Frontiers in Behavioral Neuroscience
SUMMARY

- Dopamine neurons encode a reward prediction error.
- Model-based analysis in Neurosci of Decision-making
- Reinforcement Learning models need to be refined to explain behavior / neural activity:
  - multiple parallel decision systems.
  - off-line learning during sleep.
  - meta-learning (ACC-DLPFC interactions).
- These model improvements can produce testable experimental predictions (Pavlovian autoshaping; Navigation; L-DOPA in Parkinson disease; …)
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
