dACC and the adaptive regulation of reinforcement learning parameters: neurophysiology, computational model and some robotic implementations

Mehdi Khamassi (CNRS & UPMC, Paris)

Symposim S23 chaired by Clay Holroyd on

« Formal theories of dACC function »

Reinforcement Learning (RL) framework relies on task-dpt hand-tuned parameters

$$Q(s,a) \leftarrow Q(s,a) + \alpha \,.\, \delta$$ ← **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**

Sutton & Barto (1998) MIT Press.

Meta-learning: Adaptive regulation of RL parameters

$$Q(s,a) \leftarrow Q(s,a) + \alpha \,.\, \delta$$ ← **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**

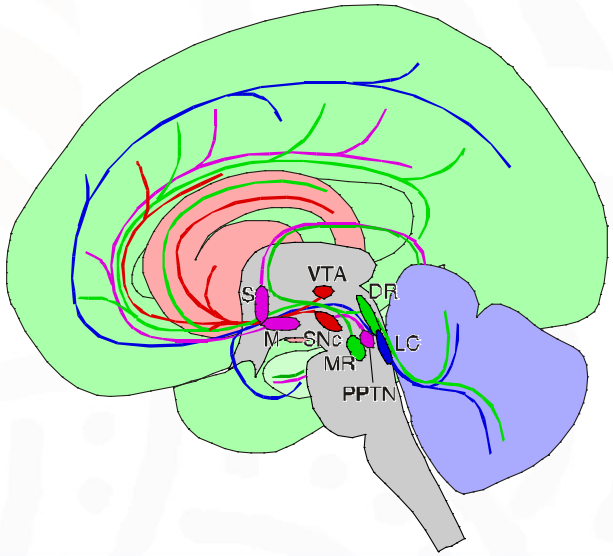
Dopamine: prediction error δ

Acetylcholine: learning rate α

Noradrenaline: exploration β

Serotonin: temporal discount γ

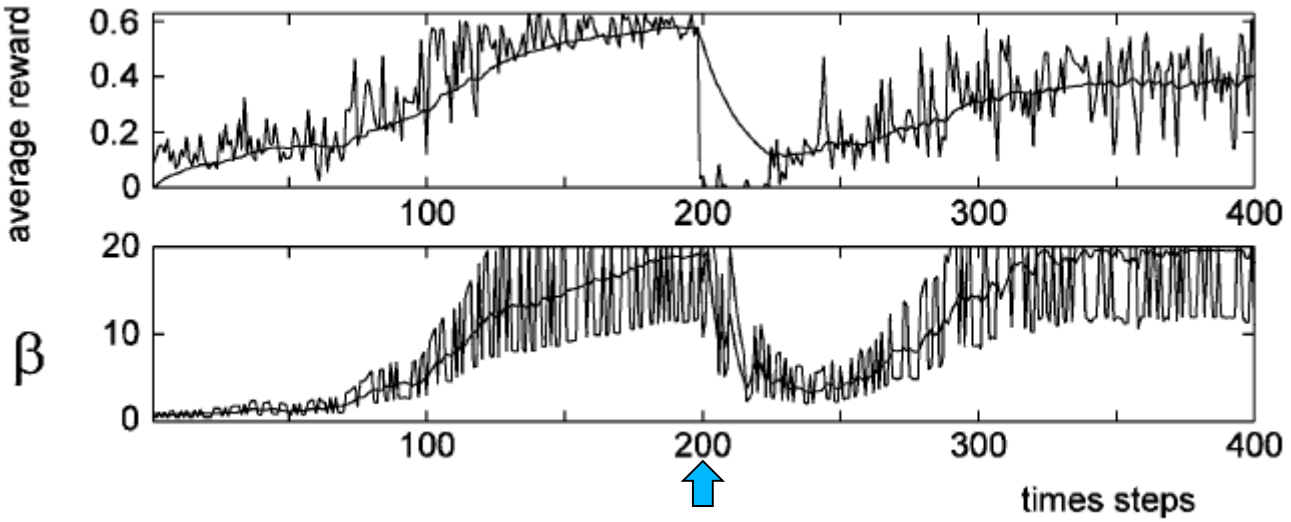
Doya (2002)



Example of meta-learning process

Meta-learning methods propose to tune RL parameters as a function of indicators of performance monitoring such as average reward and uncertainty (Schweighofer & Doya, 2003; Doya, 2008). The difference between average reward at different timescales can be used, similar to reported dACC correlates (Bernacchia et al. 2011; Wittmann et al., in the group of Matthew Rushworth).

Simulation from Schweighofer & Doya 2003

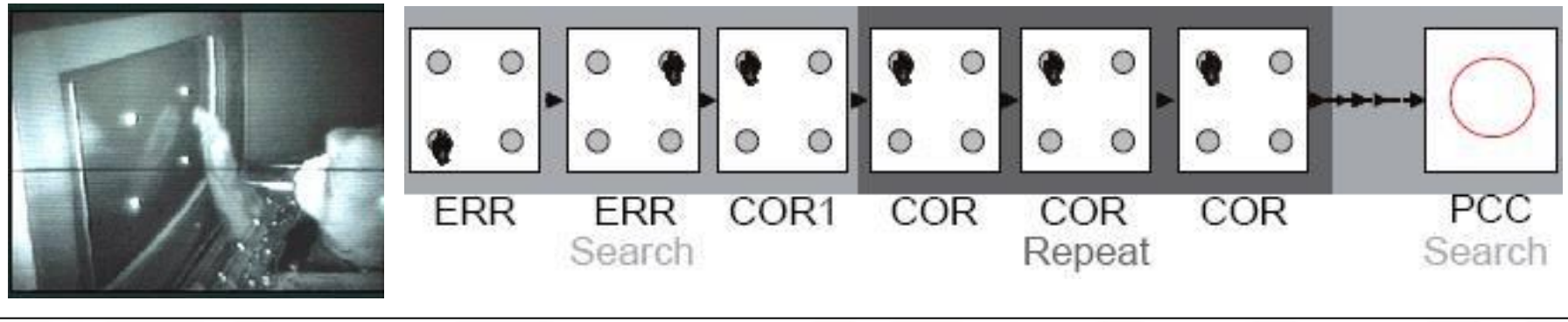


condition change
(from immediate to long-term reward)

dACC as a potential regulator of RL params based on cognitive control level

- dACC is a central component for cognitive control (Botvinick et al. 2001), an important relay of feedback signals to guide behavior (Holroyd & Coles, 2002, 2008) and to set behavioral strategies or task-sets (Dosenbach et al. 2006).
- dACC is important for feedback monitoring and show correlates of outcome history (Seo and Lee, 2007; Bernacchia et al. 2011; Wittmann et al.) and error-likelihood (Brown & Braver 2005).
- dACC could act as a regulator or energizer of decision-making processes in the LPFC (Kouneiher et al. 2009).
- dACC activity encodes volatility information, which can be used to dynamically tune the learning rate parameter in RL models (Behrens et al. 2007).

Monkey task in Emmanuel Procyk's group

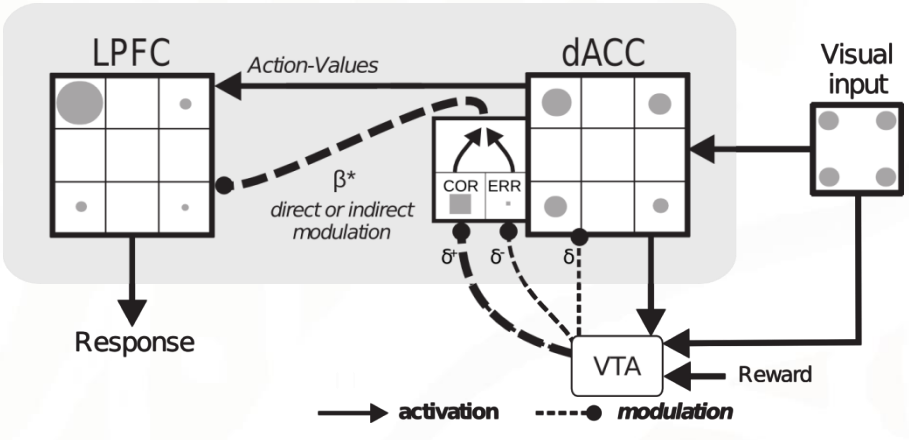


Previous results:

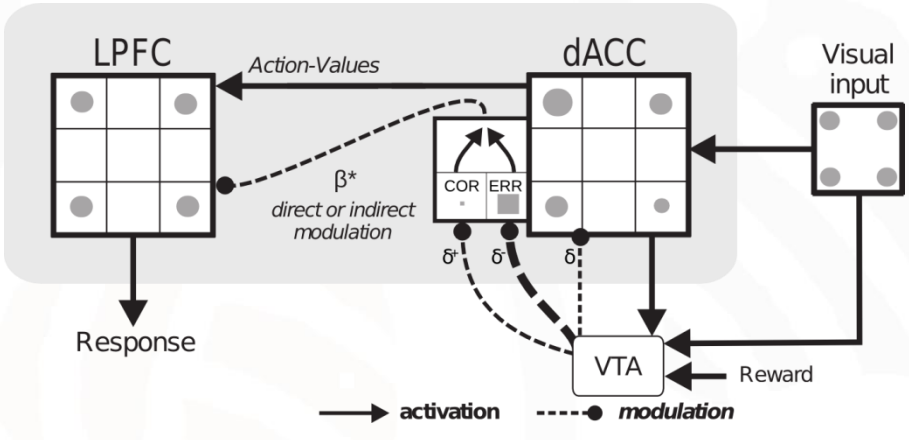
- Feedback categorization mechanisms in dACC (Quilodran et al. 2008)
- dACC neurons selective to search or repetition periods (Quilodran et al. 2008) as indicators of exploration/exploitation level?

Computational model

Series of correct trials -> exploitation



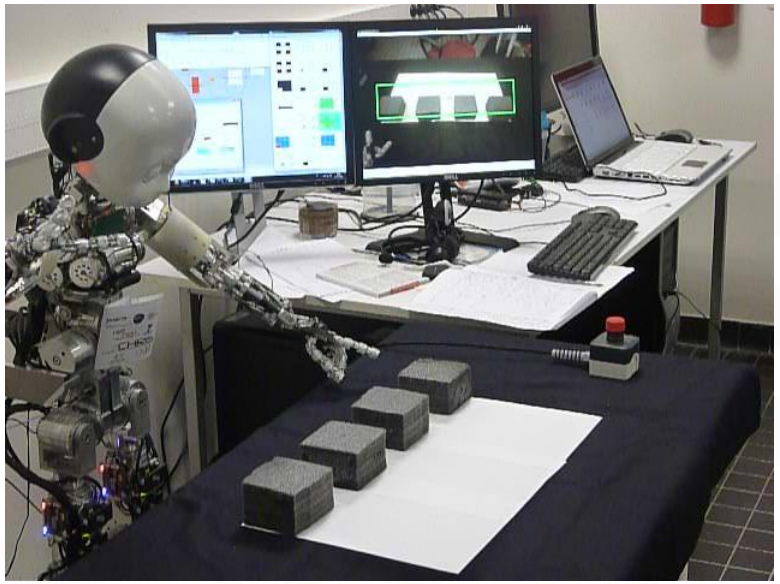
Series of error trials -> exploration



Predictions:

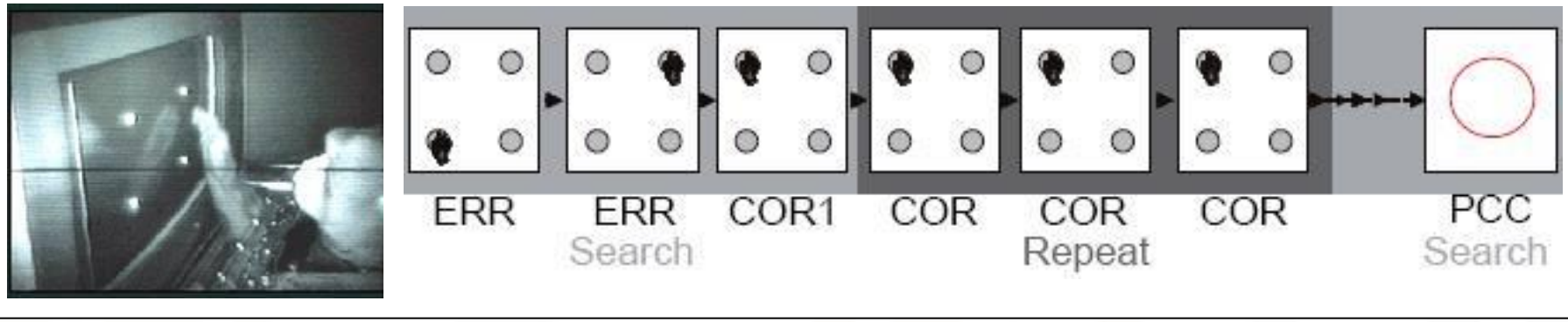
- Feedback monitoring signals in dACC (Holroyd & Coles 2002)
- Change in LPFC choice selectivity between exploration and exploitation (common predictions of LPFC exploration models; McClure et al. 2006; Krichmar 2008; Durstewitz & Seamans 2008; Cohen et al. 2007)

(Brief) sketch of robotic implementations



Reproduction of monkey performance and behavioral properties. Additional experiments to predict how monkeys may learn the task structure of Emmanuel Procyk's task (Khamassi et al. 2011 Frontiers in Neurorobotics).

New analyses of single-unit recordings



- Previous analyses of dACC feedback-related activity (Quilodran et al. 2008)
- New model-based analyses and comparison of LPFC and dACC activities in both pre-feedback and post-feedback epochs (Khamassi et al. 2014 Cerebral Cortex)

Model behavior fitting comparison

Table 1

Score obtained by each tested theoretical model, models' characteristics, and model performances to fit monkey choices for Optimization (Opt) and Test sessions

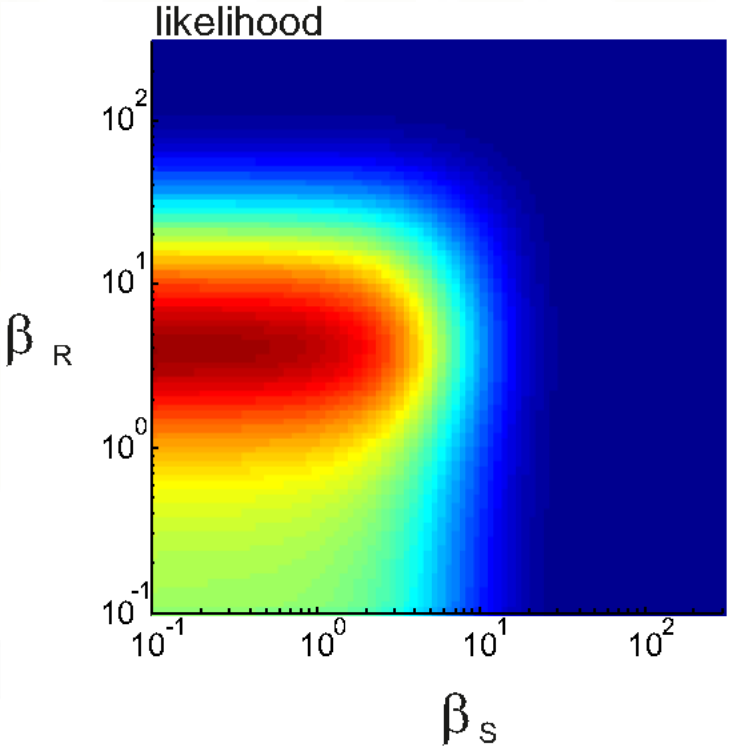
Models	r^a	RL[b]	N_P^c	Opt −LL[d]	Opt NL[e]	Opt %[f]	Opt −LPP[g]	Opt BIC/2[h]	Opt AIC/2[i]	Test −LL[d]	Test NL[e]	Test %[f]
GQLSB2β	Y	Y	4	*3290*	*0.5921*	83.47	3459	*3360*	*3298*	29732	0.5830	*74.17*
SBnoA2β	Y	N	3	3385	0.5831	84.13	*3422*	3438	3391	30901	0.5708	73.11
GQLSB	Y	Y	3	3355	0.5859	83.80	3502	3408	3361	*29539*	*0.5850*	73.43
SBnoA	Y	N	2	3454	0.5768	84.29	3480	3489	3458	30613	0.5738	72.59
SBnoF	Y	N	1	3586	0.5648	*84.43*	3604	3604	3588	32169	0.5578	71.61
GQLBnoS	Y	Y	3	3721	0.5528	78.59	3847	3773	3727	33274	0.5467	69.47
GQLSnoB	Y	Y	3	3712	0.5536	76.66	3843	3764	3718	31501	0.5646	70.12
GQLnoSnoB	Y	Y	3	4253	0.5079	69.14	4292	4305	4259	35376	0.5262	66.60
GQL	N	Y	3	5590	0.4104	65.10	5994	5643	5596	49282	0.4089	53.20
QL	N	Y	2	5960	0.3869	44.92	7755	5995	5964	59734	0.3382	48.78
ClockS	Y	N	2	5249	0.4333	70.92	5841	5284	5253	47504	0.4223	58.71
RandS	Y	N	1	4607	0.4800	69.43	4621	4624	4609	39488	0.4884	63.73

Classical reinforcement learning models (QL/GQL) and control logical models (ClockS/RandS) cannot reproduce monkey behavior. Models combining RL and task structure information (GQLSB2β;SBnoA2β;GQLSB) can.

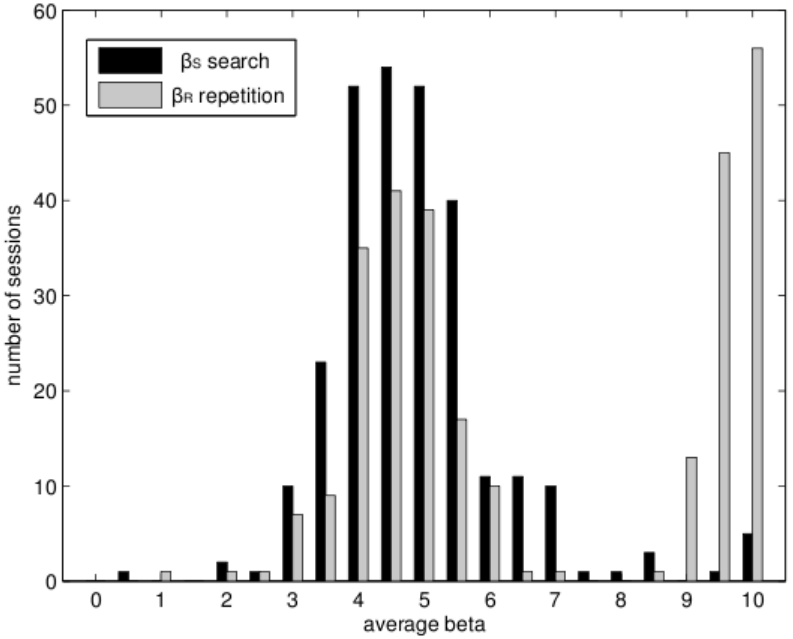
Khamassi et al. 2014 Cerebral Cortex

Behavioral shifts found by difference in model's optimized exploration parameters

Example session

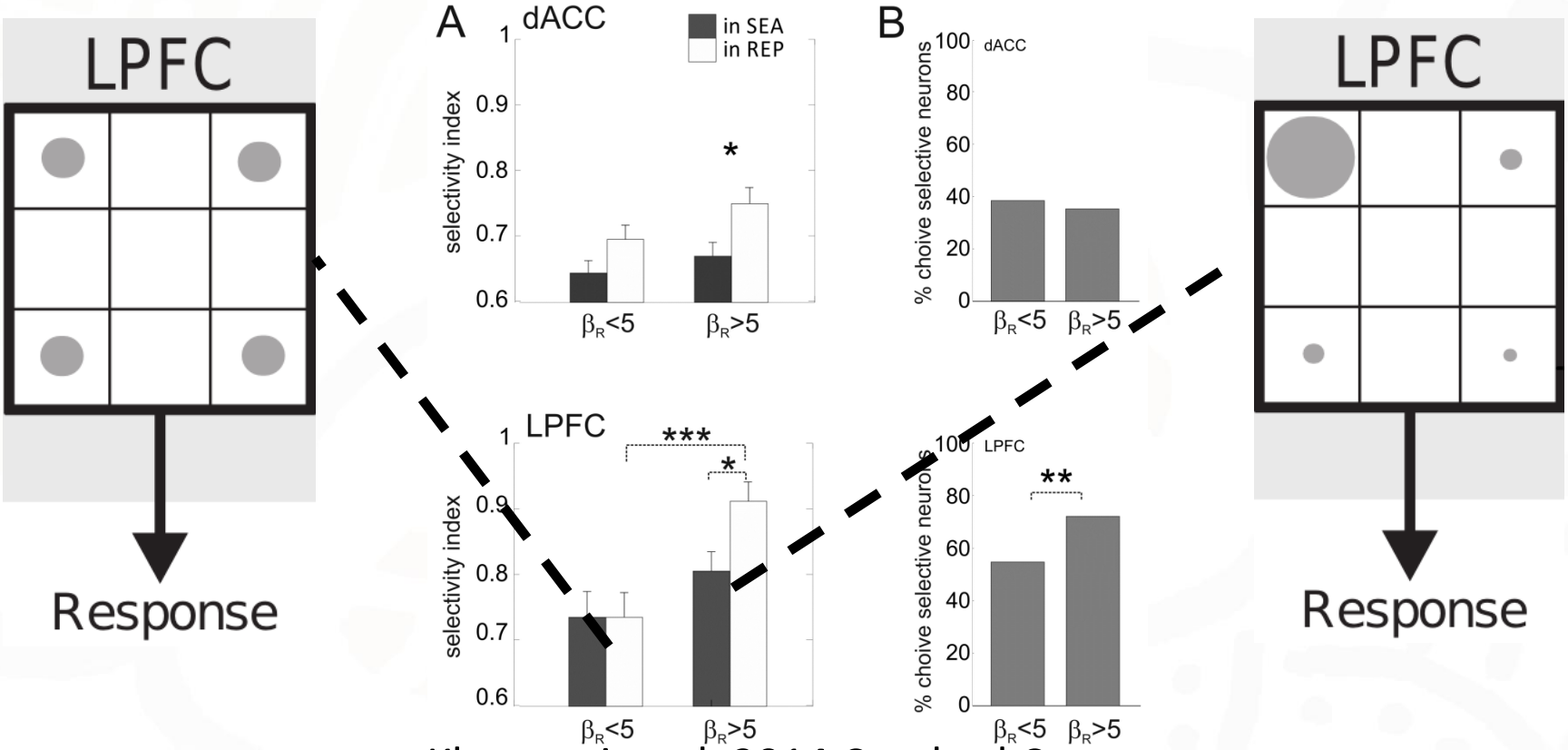


All sessions



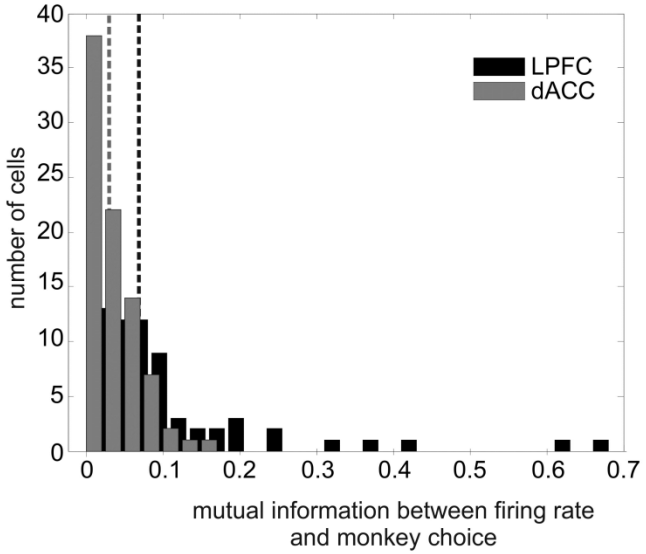
Khamassi et al. 2014 Cerebral Cortex

Increases in LPFC choice selectivity follow changes in the model's optimized exploration parameter



Khamassi et al. 2014 Cerebral Cortex

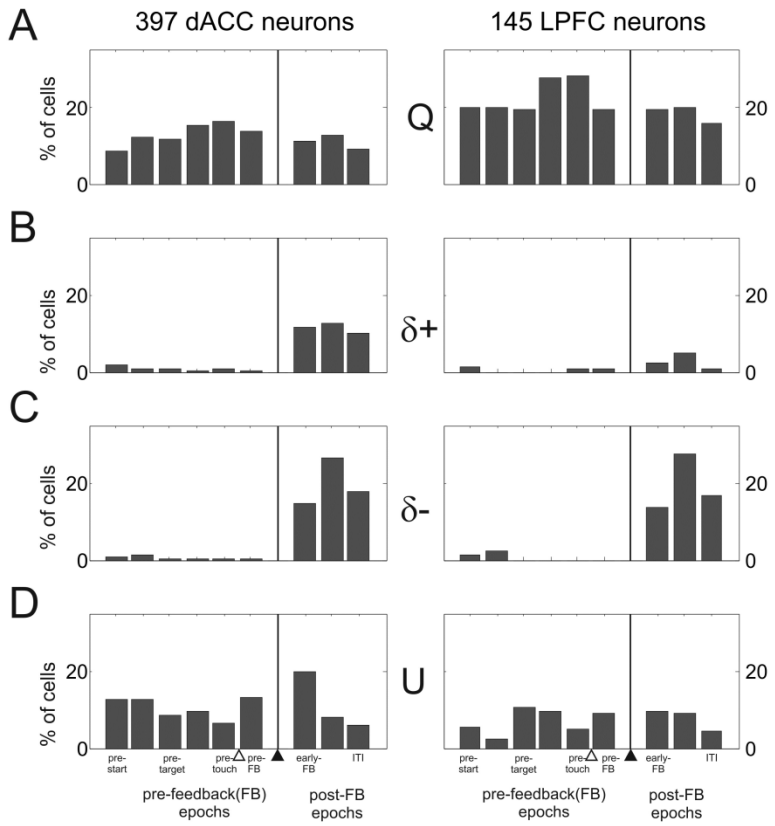
Higher mutual information between monkey choice and LPFC activity than with dACC activity

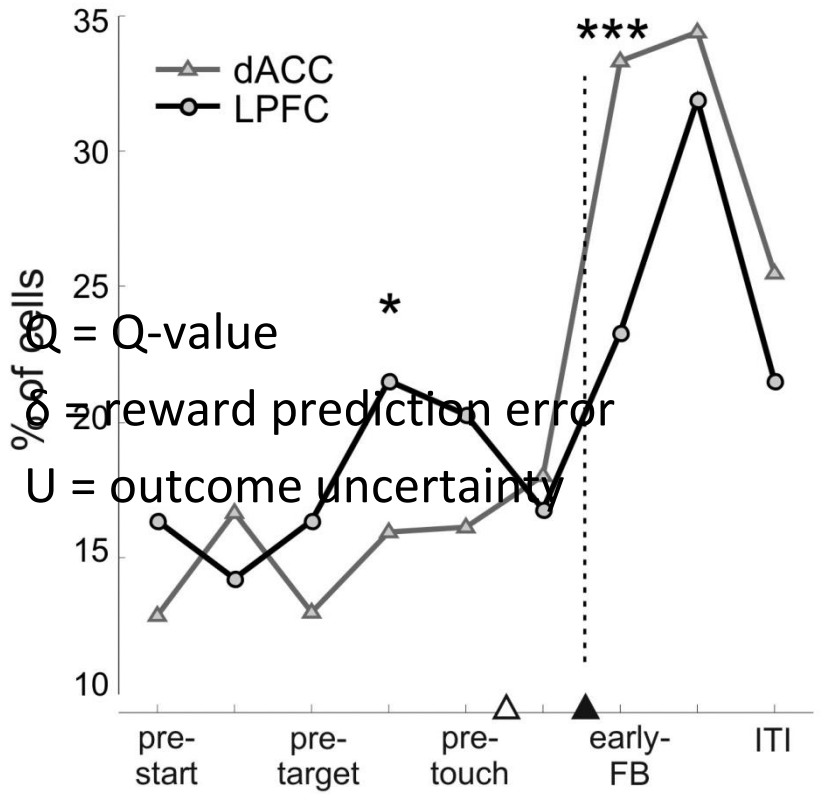


Subset of LPFC neurons with high mutual information (MI) with choice.

Khamassi et al. 2014 Cerebral Cortex

Single-unit correlates of model variables





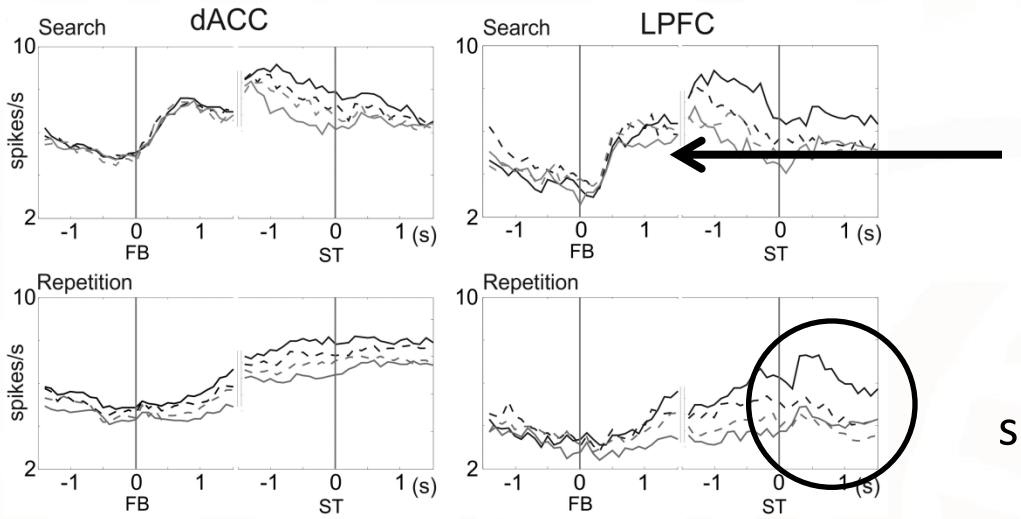
Q = Q-value

δ = reward prediction error

U = outcome uncertainty

Khamassi et al. 2014 Cerebral Cortex

Multiplexing in single-unit activity



Response to errors during search

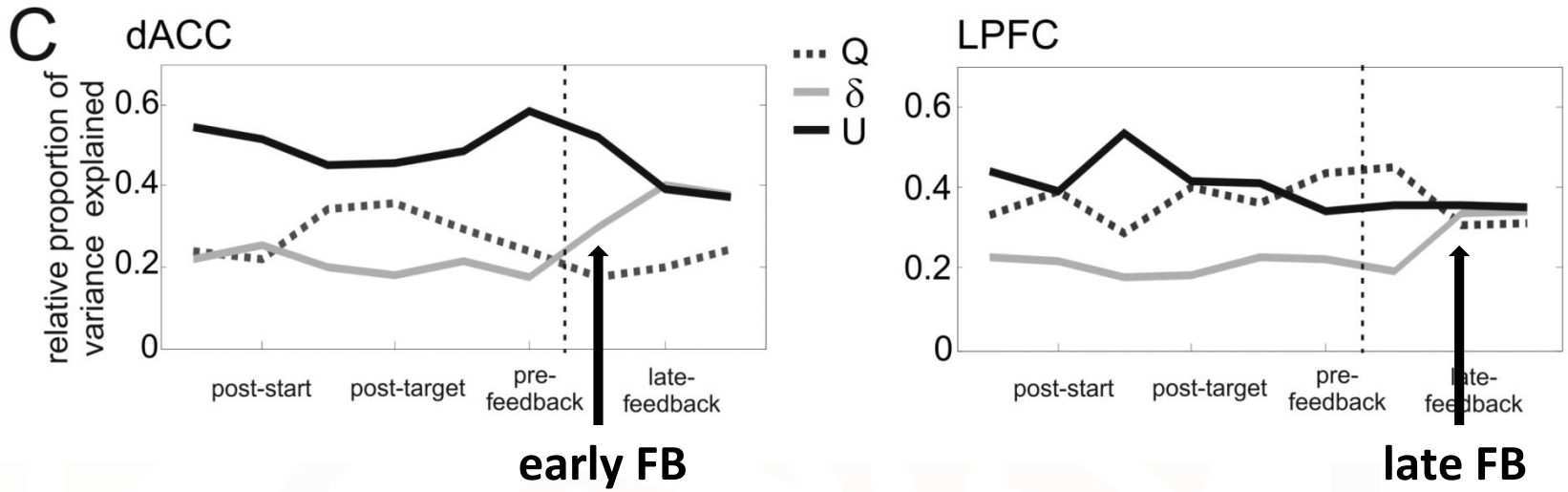
Increase in choice selectivity during repetition

LPFC neurons with activity correlated with **negative prediction error** at the time of the feedback show an **increase in delay choice selectivity** at the next trial.

Khamassi et al. 2014 Cerebral Cortex

Contribution to activity variance

PCA Analysis on regression coefficient of model variables



Before feedback, outcome uncertainty (U) dominates in dACC, while more integrated with Q-values in LPFC.

After feedback, earlier correlates of prediction error in dACC than in LPFC.

Khamassi et al. 2014 Cerebral Cortex

Summary

- dACC is in an appropriate position to monitore feedback and to modulate learning parameters that could influence decision-making in LPFC (meta-learning).
- LPFC single-unit activity was more tightly related to monkey choices, and choice selectivity varied according to changes in the exploration parameter.
- dACC activity more dominantly tracked outcome uncertainty before the feedback, and prediction errors in the early-feedback epoch.
- Such a pluridisciplinary approach can contribute both to a better understanding of the brain and to the design of algorithms for autonomous decision-making in robots.

Acknowledgments

Neurophysiology team, INSERM, Lyon

Emmanuel Procyk
René Quilodran
Charlie R. Wilson

Robotics team, INSERM, Lyon

Peter F. Dominey
Pierre Enel
Stéphane Lallée

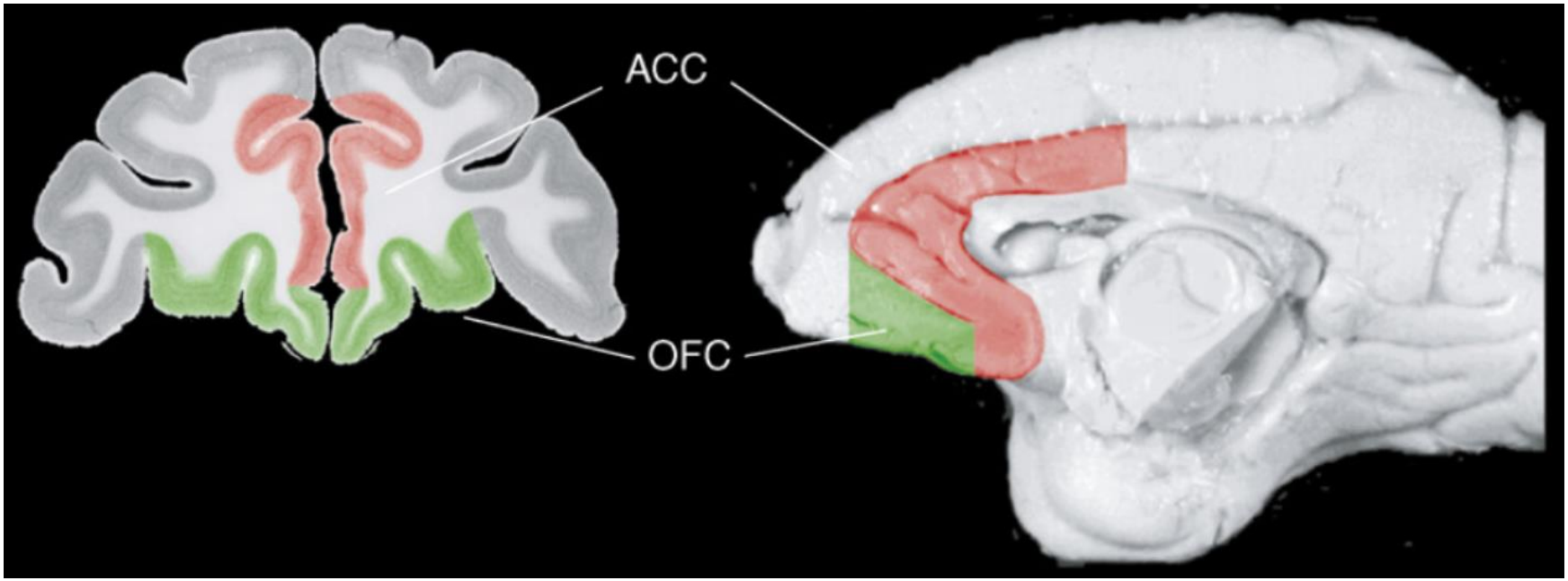
Financial support

French National Research Agency (ANR) Learning under Uncertainty Project and AMORCES Project.

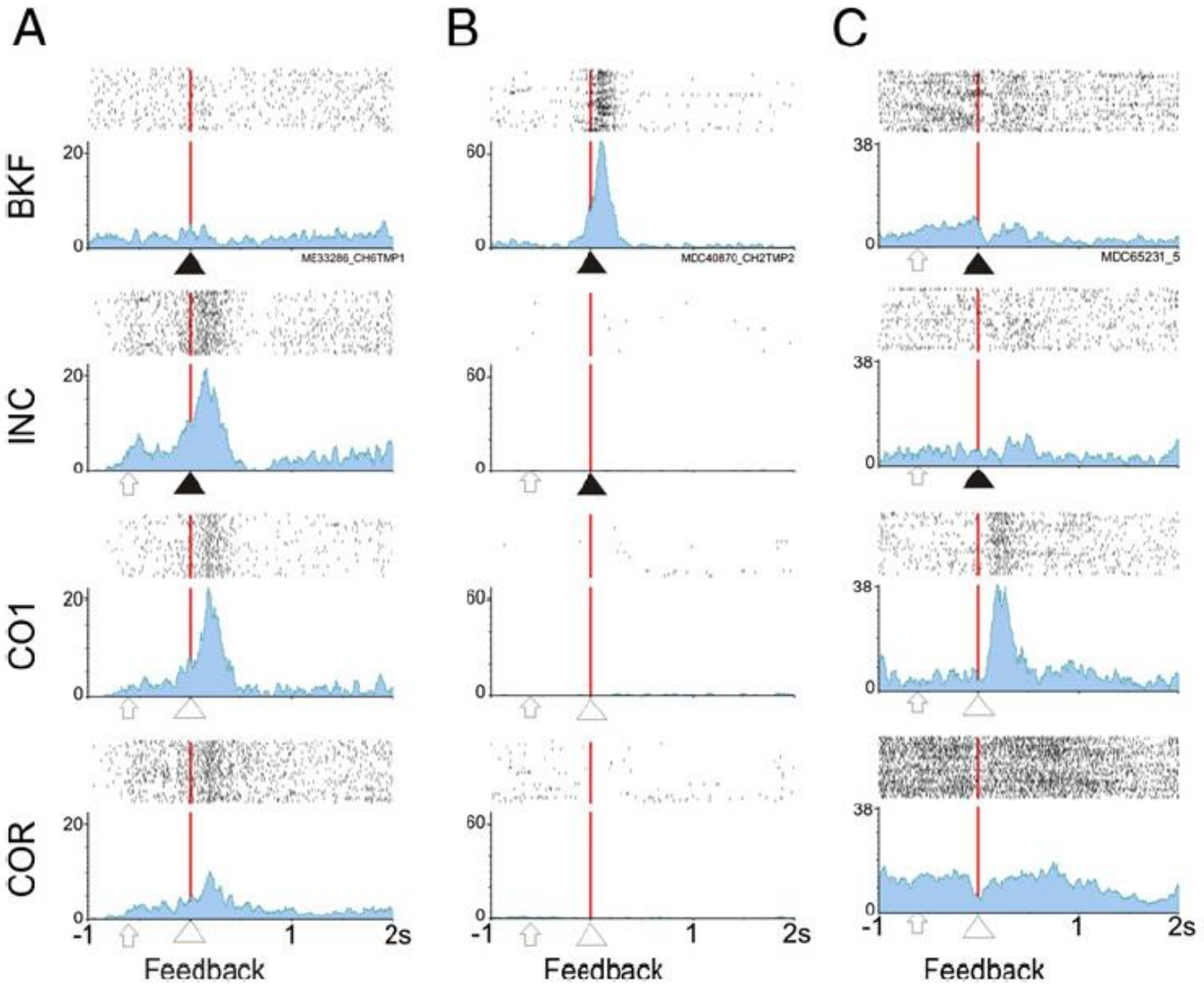
FINANCÉ PAR ANR

MEDIAL AND LATERAL PREFONTAL CORTEX

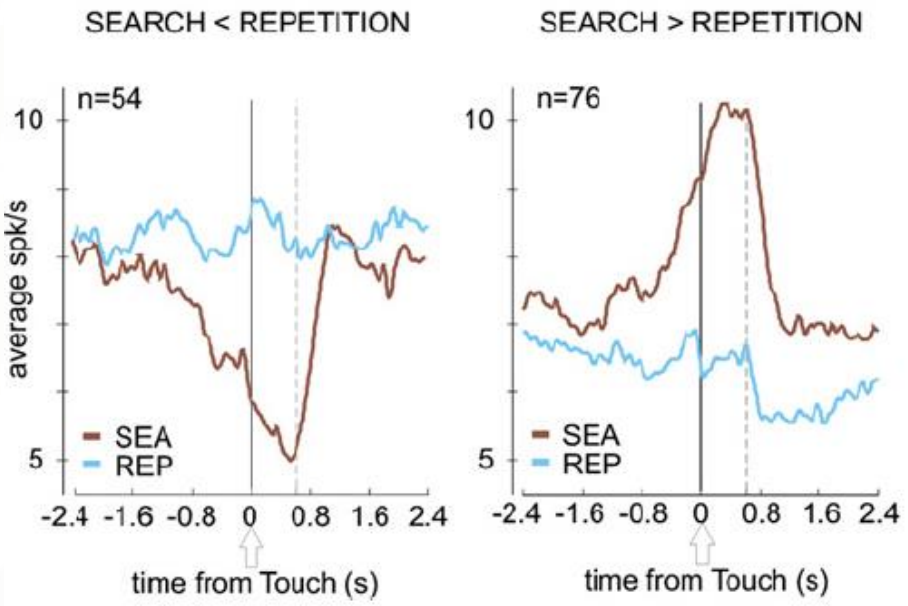
- Anatomy figure with ACC and LPFC.



Frontal cortex activity related to exploration/exploitation



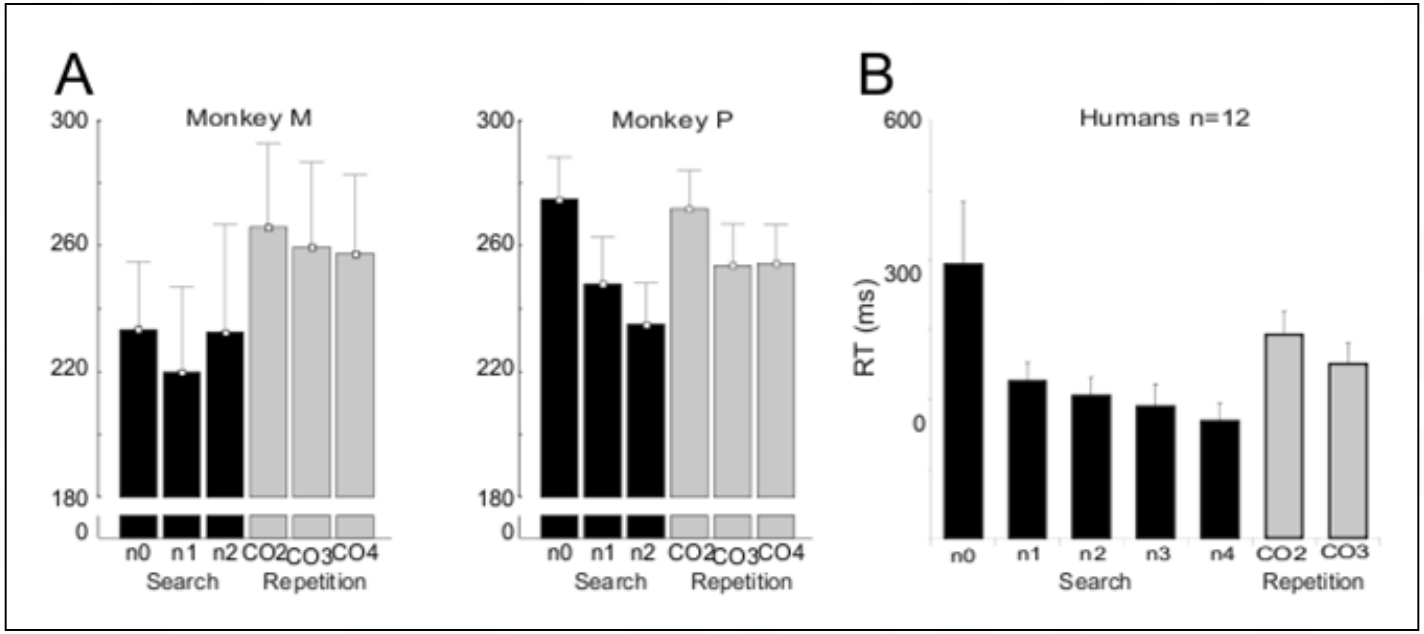
PREVIOUS RESULTS



- Different ACC subpopulations of neurons are selective to SEARCH and REPETITION periods.

(Quilodran et al., 2008)

Model-based analysis of monkey behavioral data

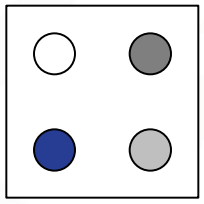


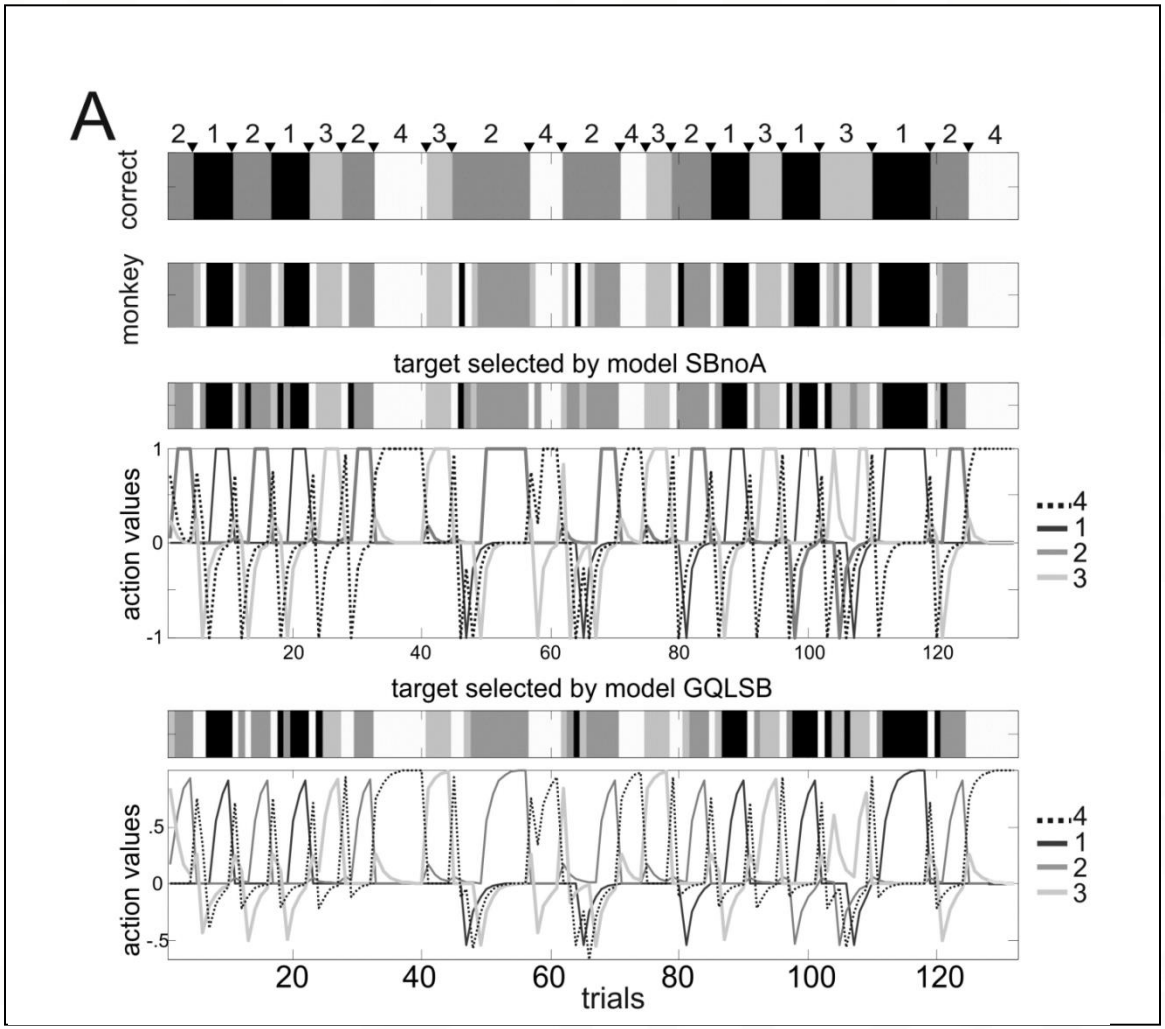
Reaction times

Model-based analysis of monkey behavioral data

Model-based analysis of behavior

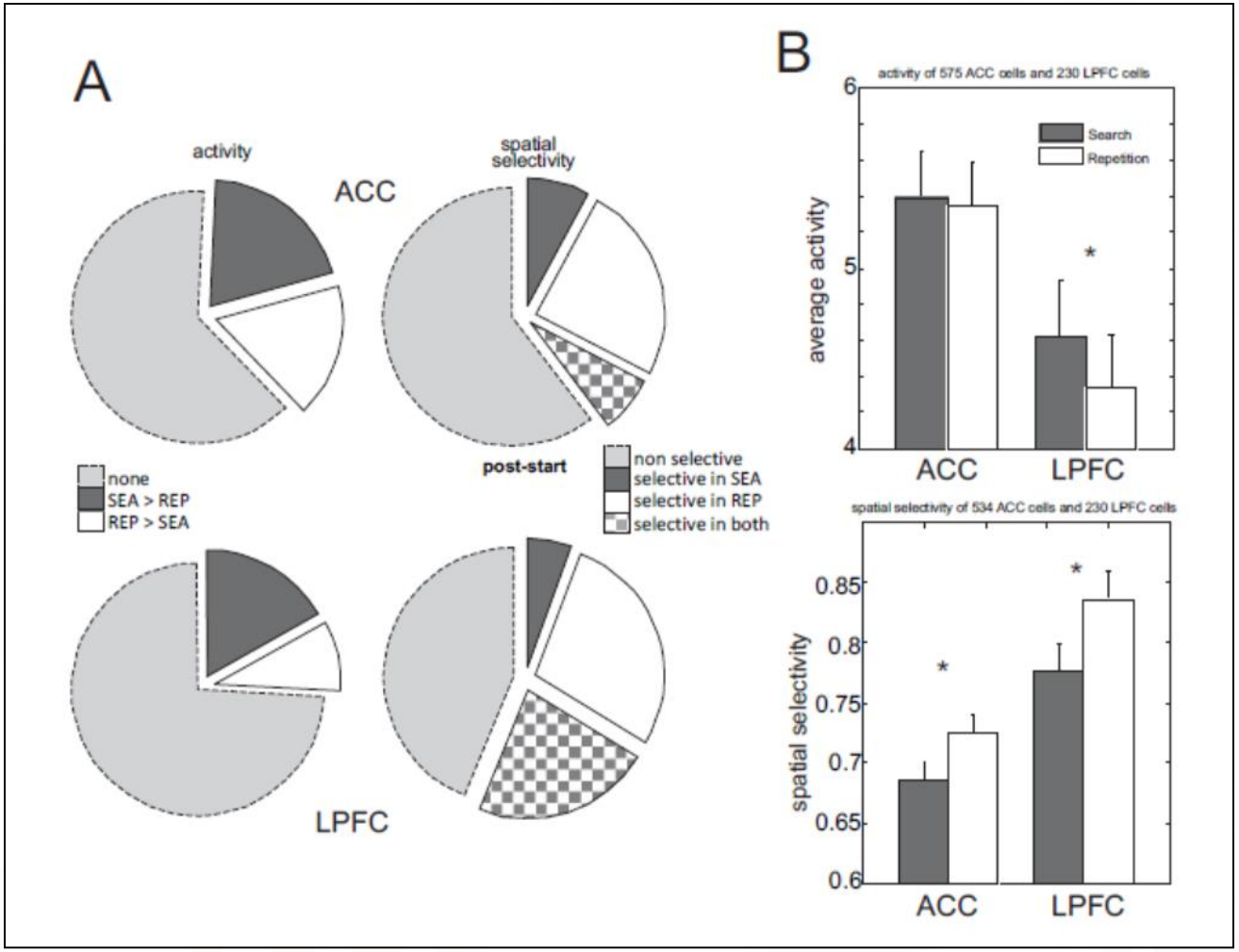
75% similarity (likelihood=0.6)





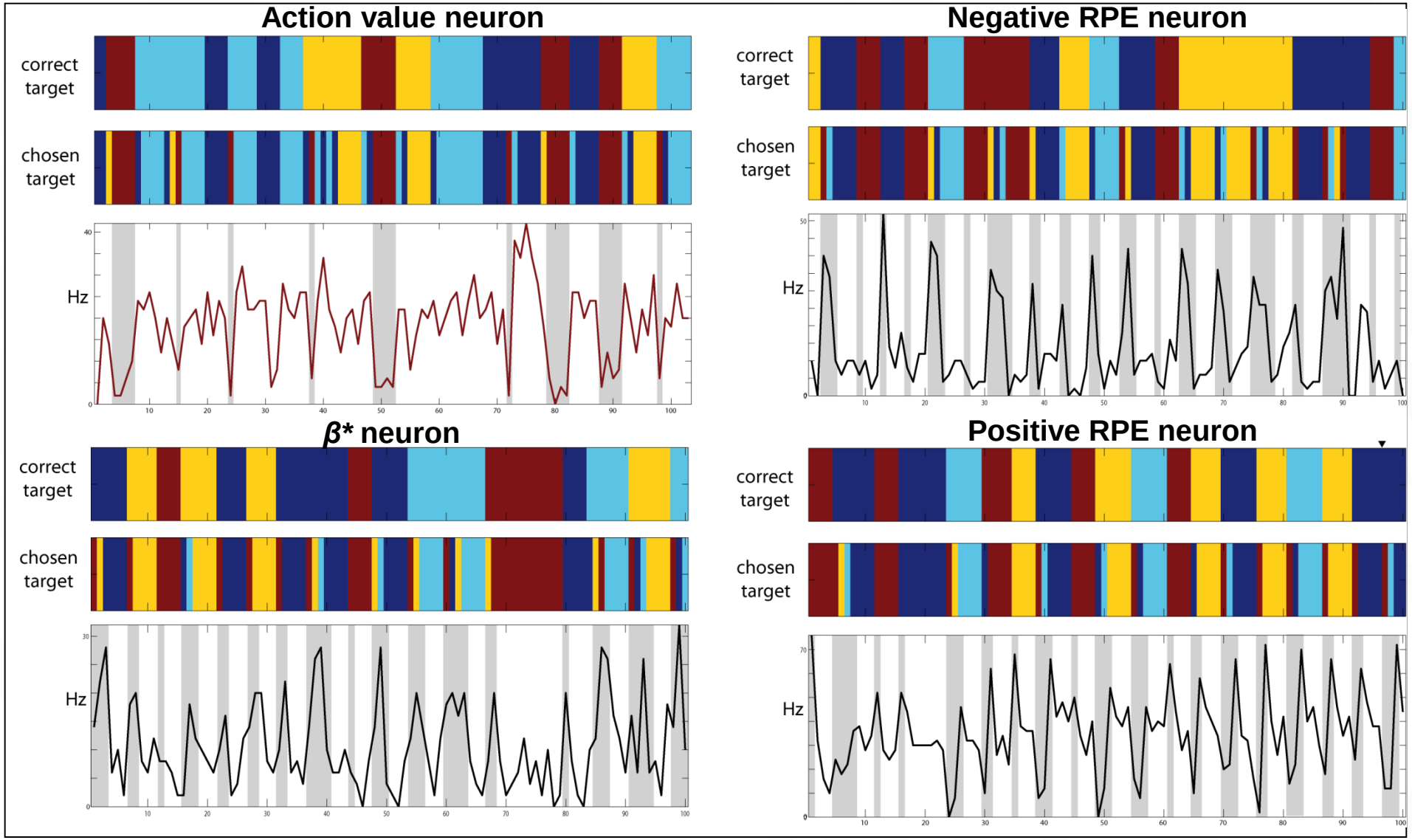
Khamassi et al. (2014) Cerebral Cortex

Activity variation between SEARCH and REPEAT periods



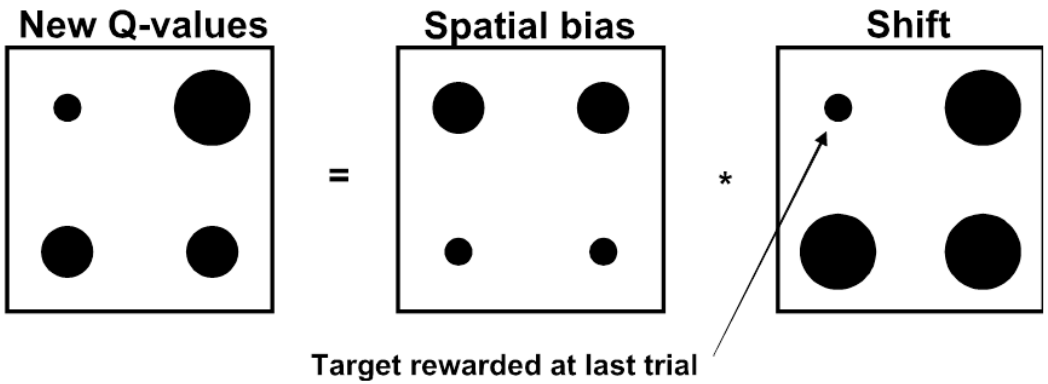
Global decrease of activity during the repetition period, and increase in spatial selectivity, as predicted by *β in the model**

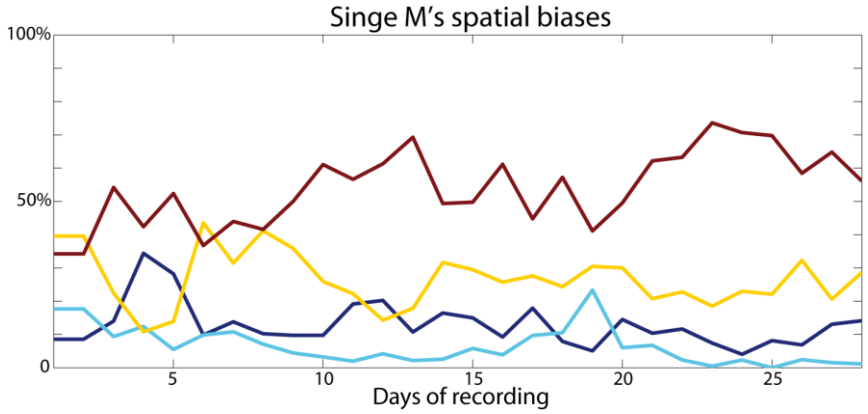
Model-based analysis of neuronal data

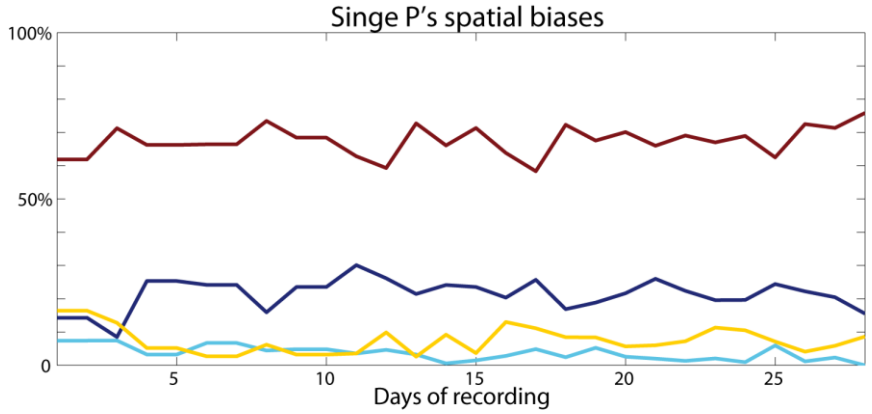


Predicting monkeys' choices

The RL model needs a reset mechanism at the beginning of each new problem to reproduce monkey behavior = **knowledge about task structure**







Model-based analysis of neuronal data

Negative RPE neuron

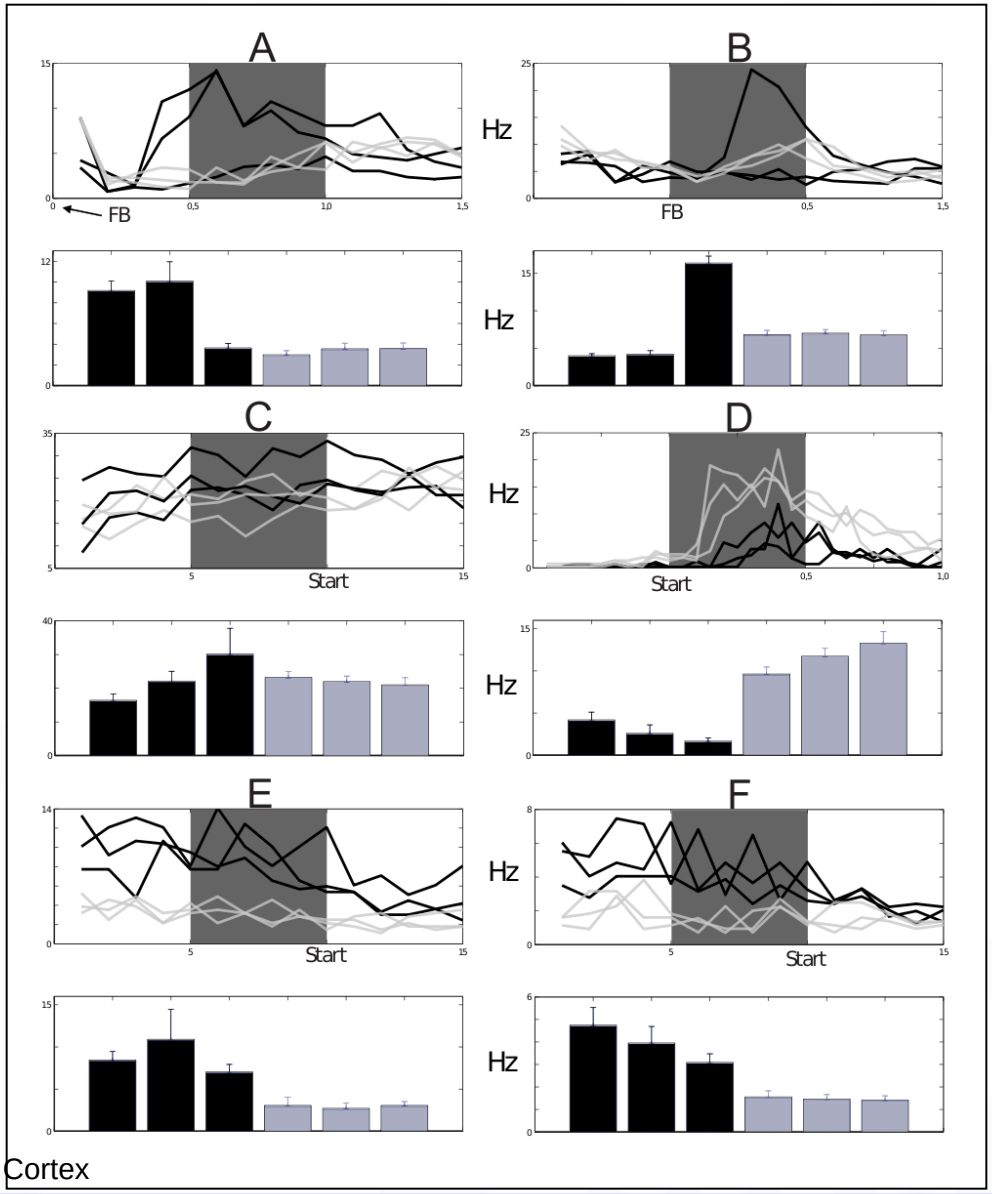
Positive RPE neuron

***U* neuron**

Opposite *U* neuron

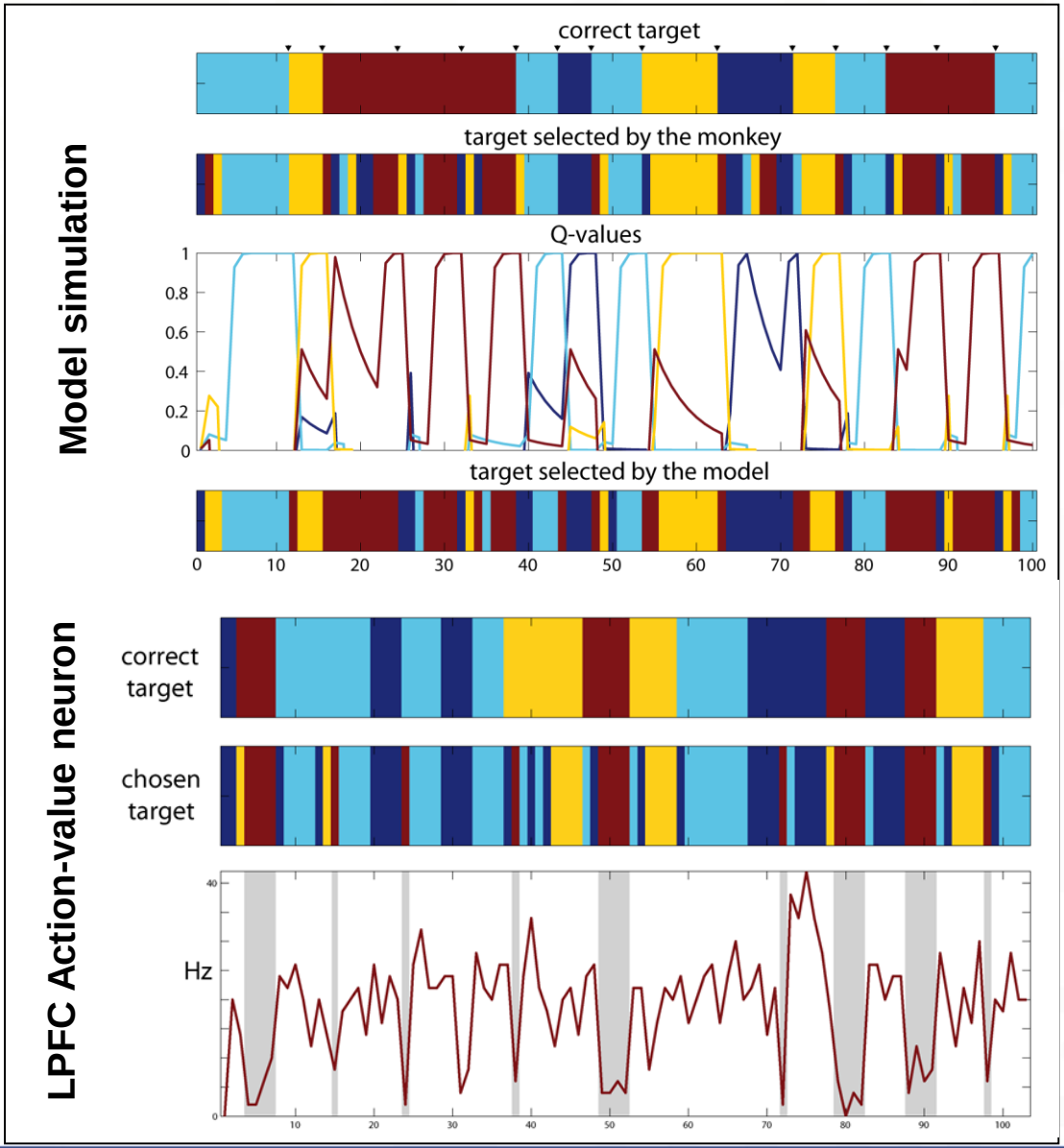
SEARCH/REPEAT neuron

SEARCH/REPEAT neuron



Khamassi et al. (2014) Cerebral Cortex

Model-based analysis of neuronal data

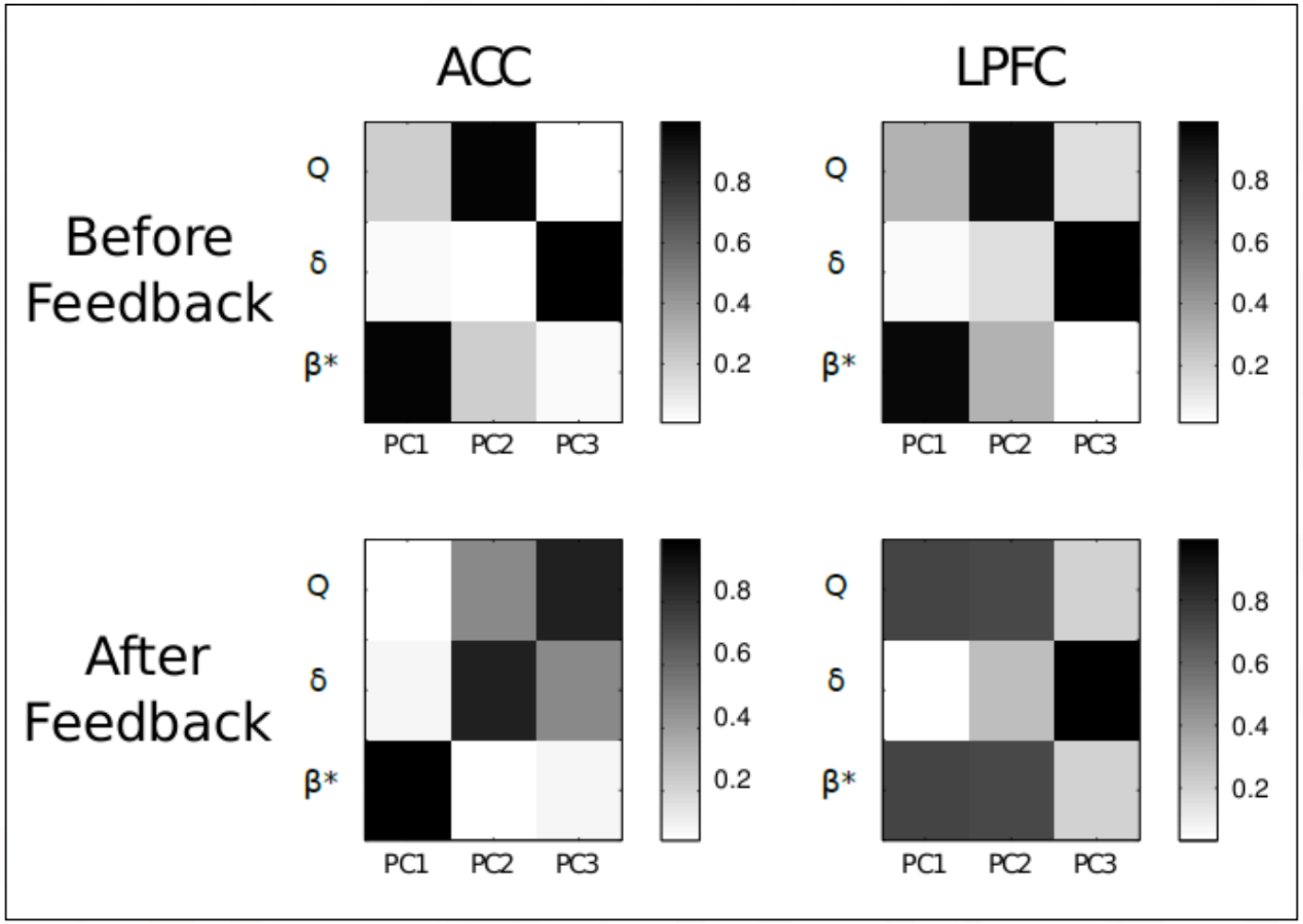


Integration of different model variables according to PCA analysis

neurons' firing rate

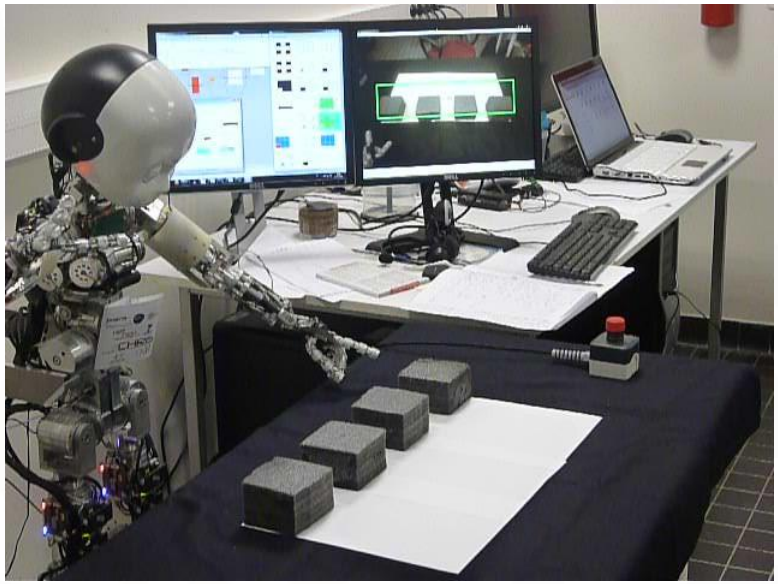
f1=**a***Q4+**b***RPE+**c***MV+..
f2=**d***Q4+**e***RPE+**f***MV+..

Principal Component Analysis (PCA)



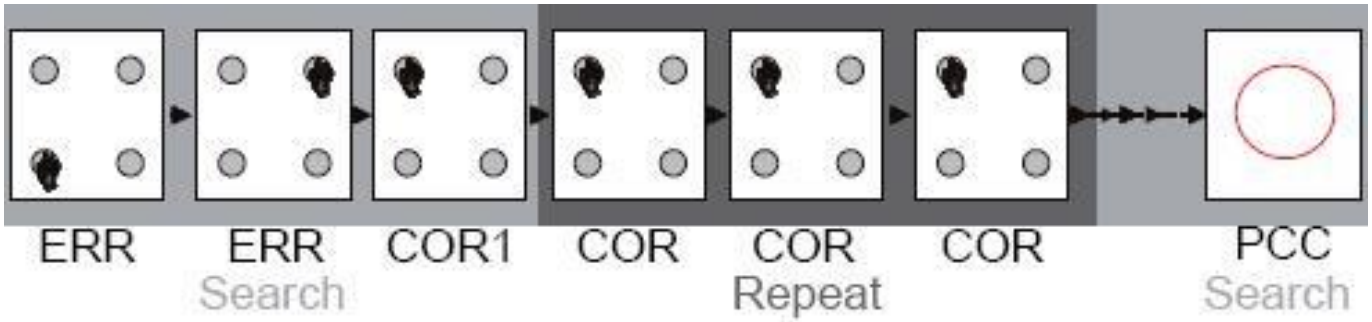
β* **is more integrated with action values in LPFC than in ACC**

(Brief) sketch of robotic implementations



Reproduction of monkey performance and behavioral properties. Additional experiments to predict how monkeys learn the task structure of Emmanuel Procyk's task (Khamassi et al. 2011 Frontiers in Neurorobotics).

Meta-learning applied to Human-Robot Interaction

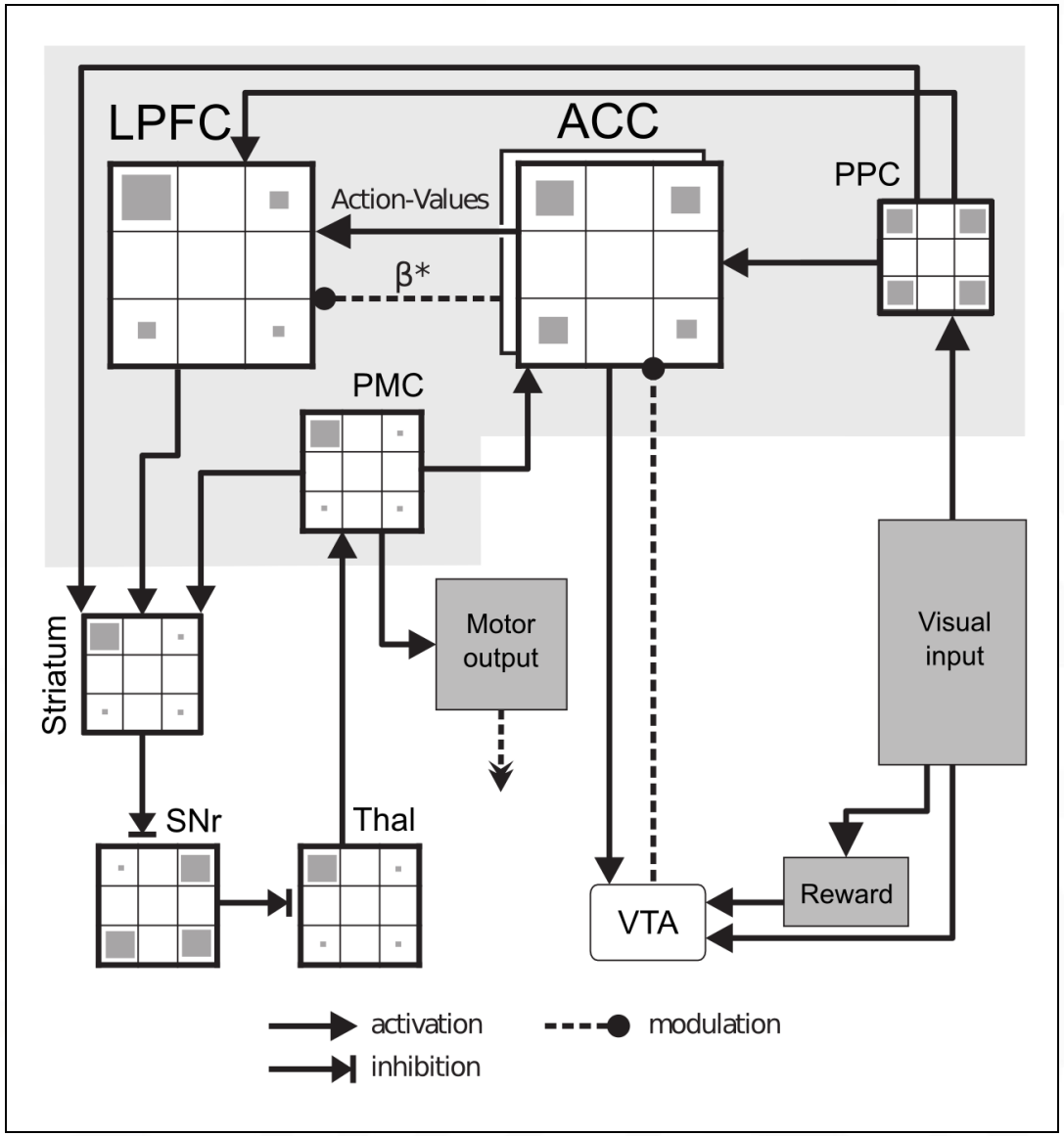


- In the previous task, monkeys and the model a priori 'know' that *PCC* means a reset of exploration rate and action values.
- Here, we want the iCub robot to learn it by itself.

Computational model

β*: feedback history

used to tune β



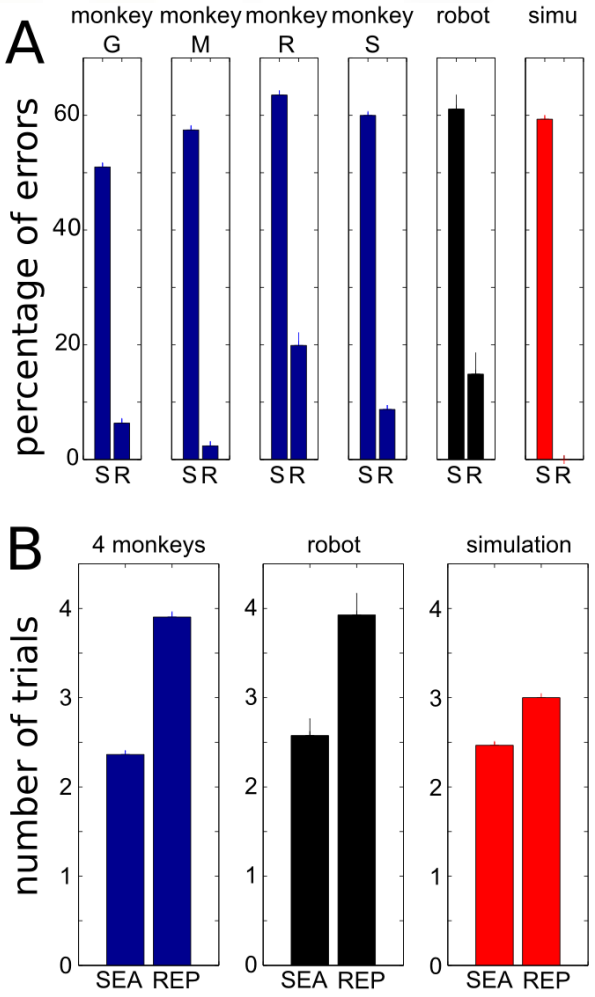
Robotic arm

video

Robotic arm

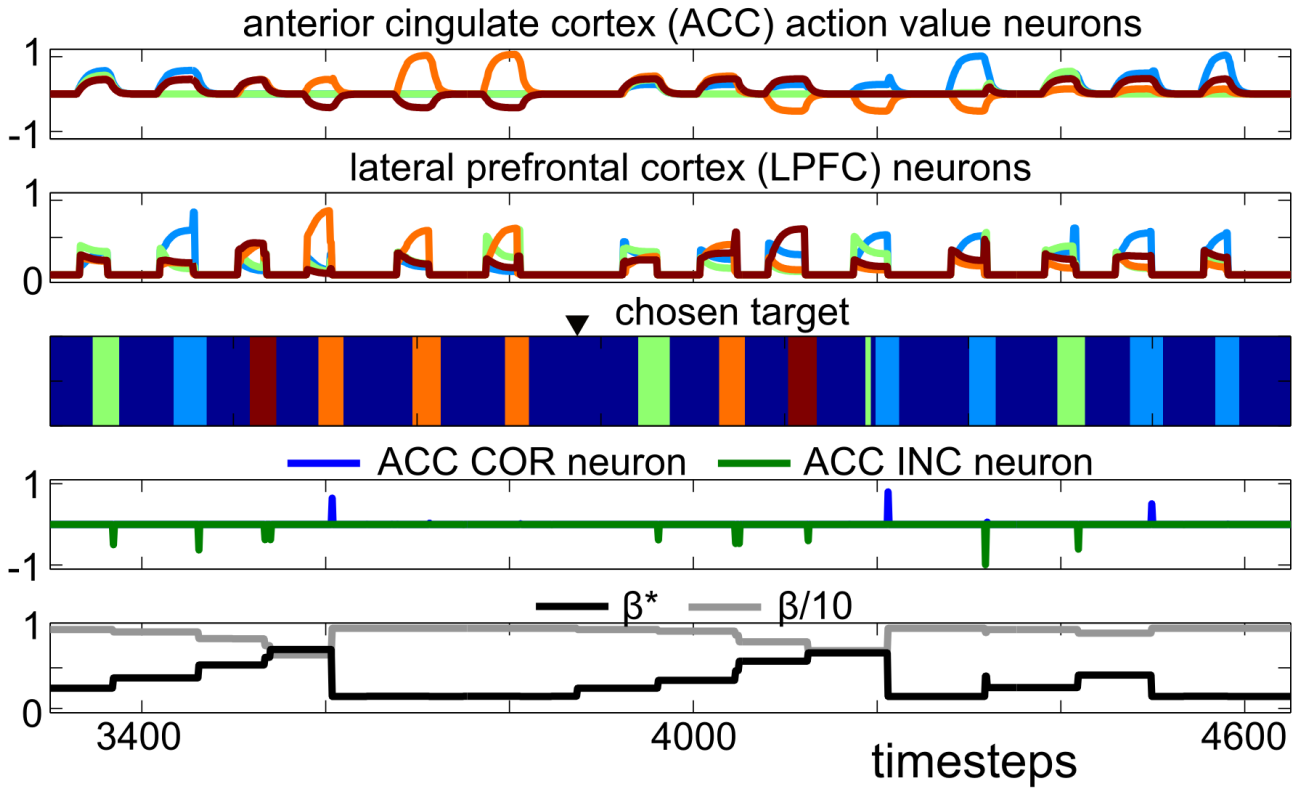
Computational model simulation

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



Meta-learning applied to Human-Robot Interaction

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

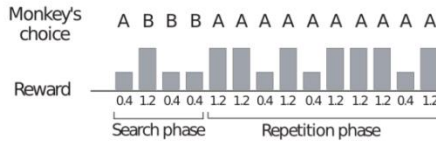


Simulation on the model on a probabilistic task (Amiez, 2006)

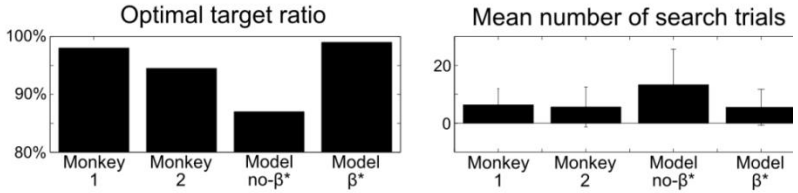
A

Amount of juice dispensed	Target A	Target B
1.2 mL	70%	30%
0.4 mL	30%	70%

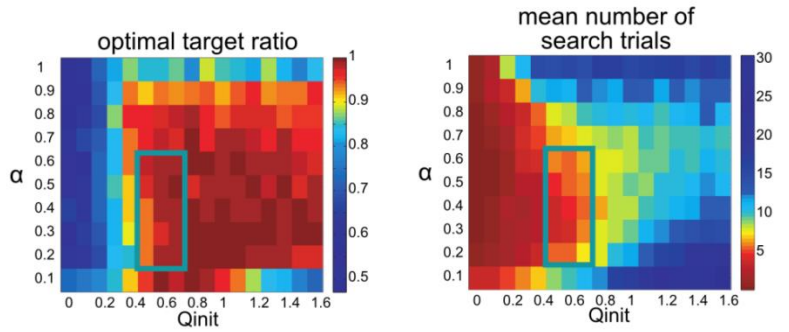
Stochastic task with 2 targets



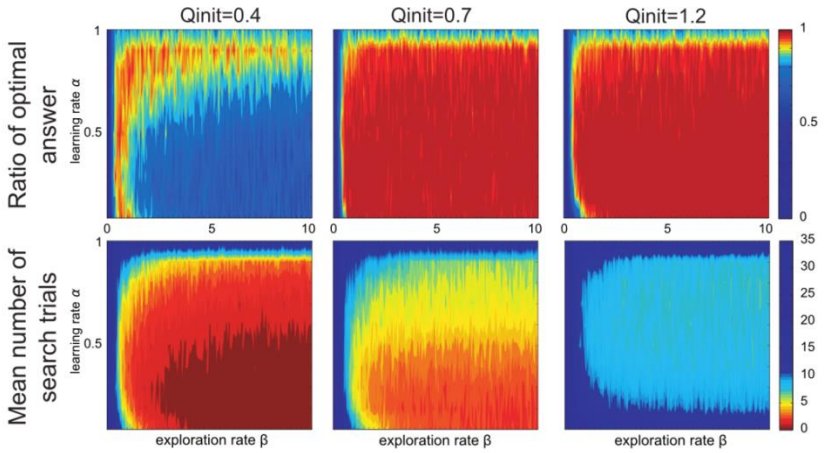
B



C



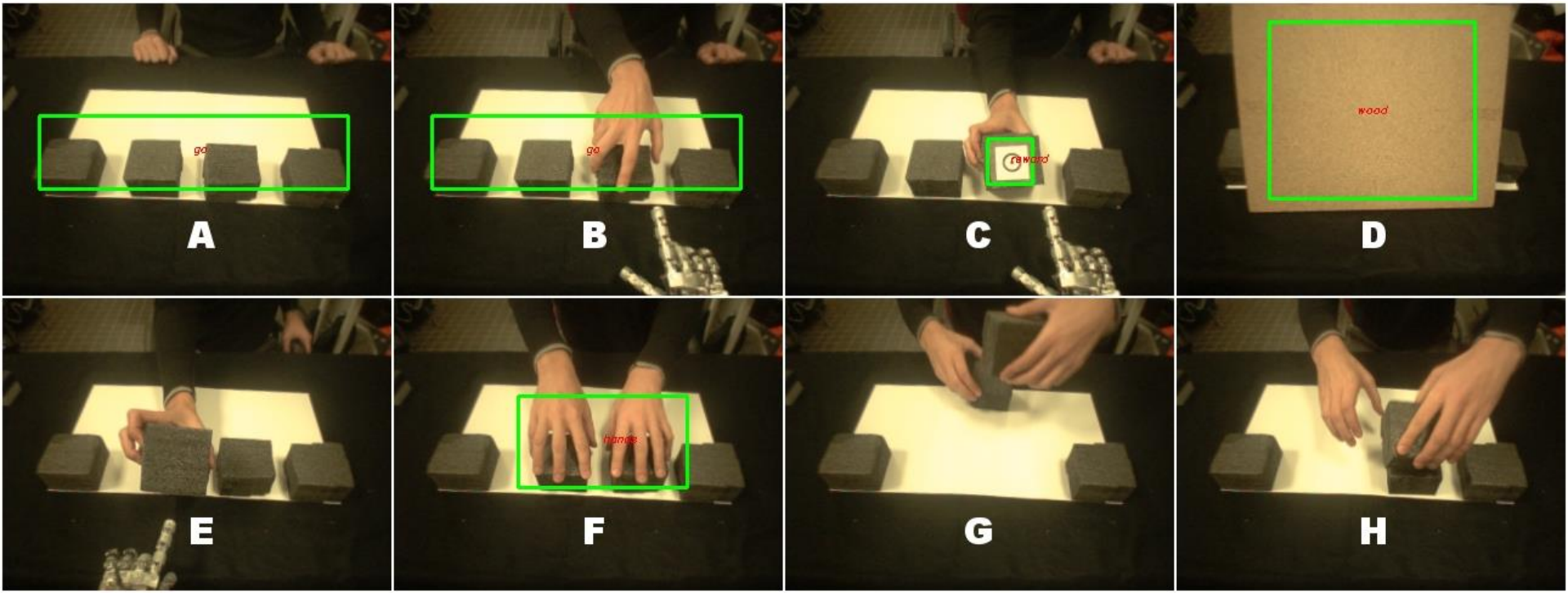
D



Meta-learning applied to Human-Robot Interaction

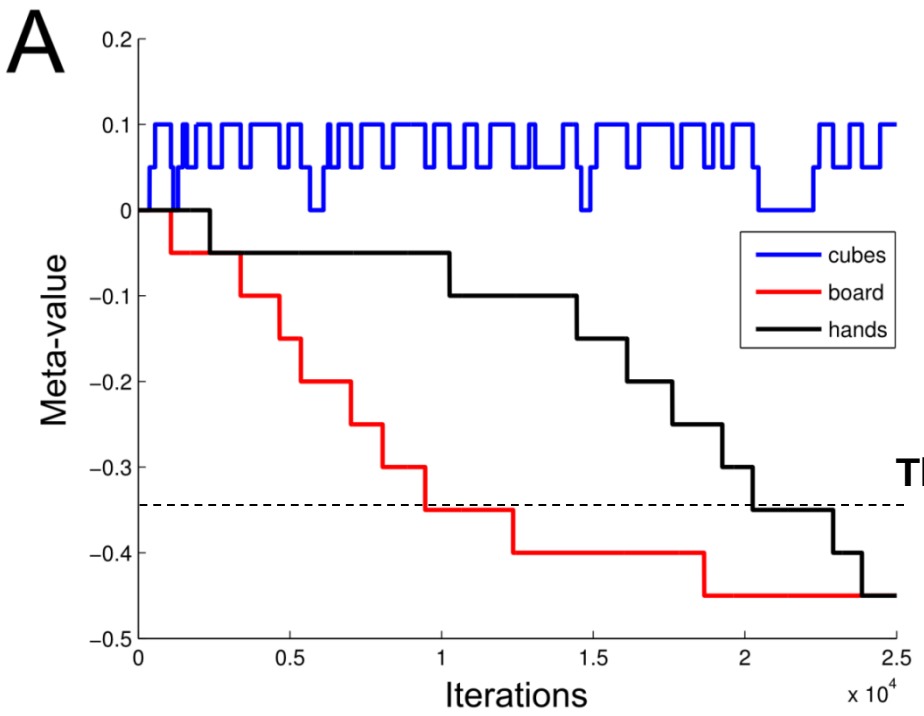
Khamassi et al. (2011) Frontiers in Neurorobotics

Meta-learning applied to Human-Robot Interaction

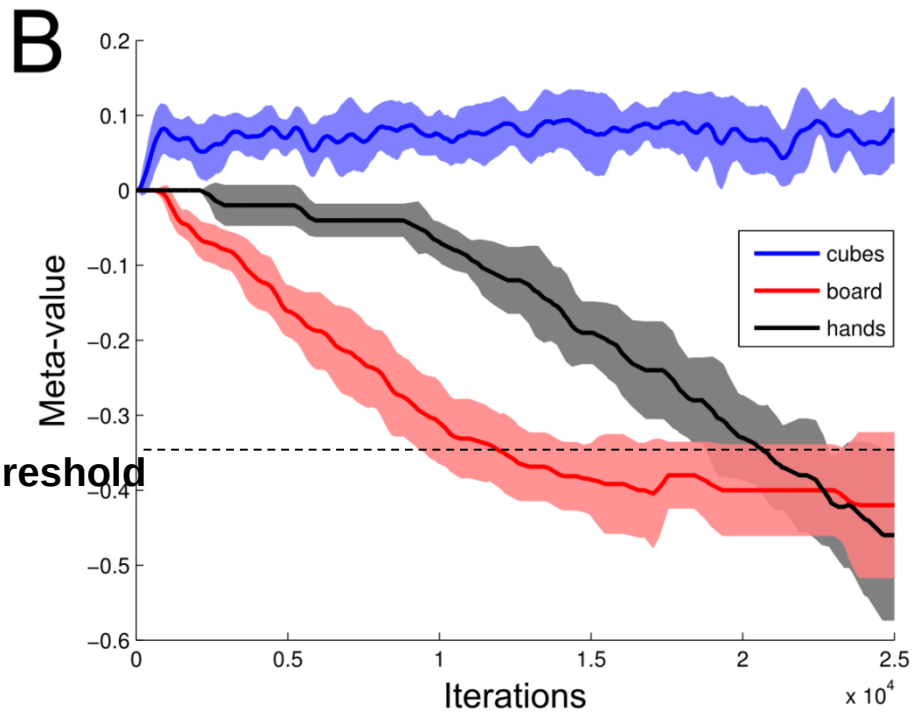


Meta-learning applied to Human-Robot Interaction

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

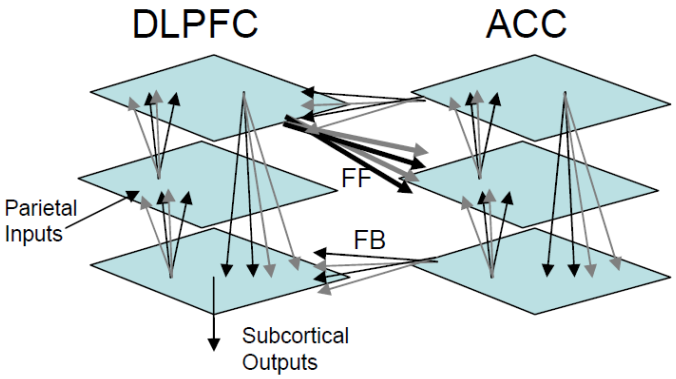


Threshold



PERSPECTIVES

- Approche modélisation utile pour formaliser plus précisément la fonction de groupes de neurones
- Notre modèle propose un mécanisme pour intégrer RL et TM dans ACC
- Pour aller plus loin, il faudrait modéliser cette dynamique de populations avec recurrent neural nets



Cortical laminar structure and Cx-Cx connectivity