Learning to make reward-guided decisions: sequential, successive, and social

Hiro Nakahara

NAKA HARA LAB

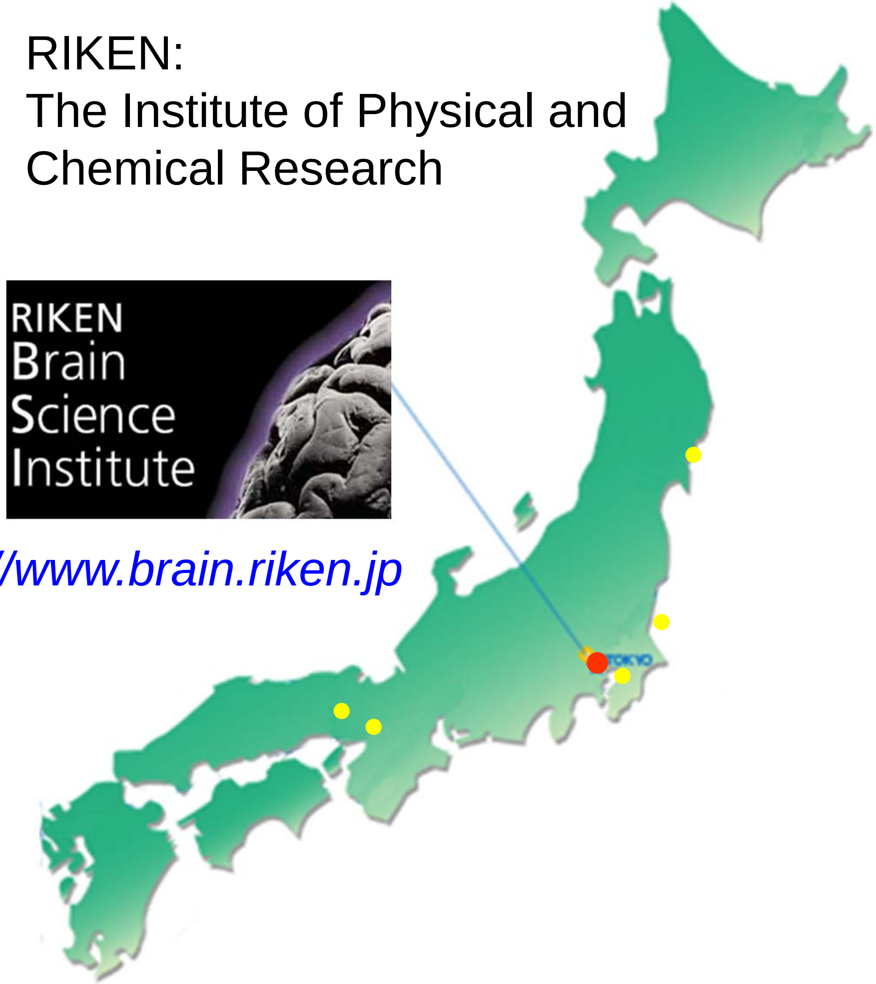
Lab for Integrated Theoretical Neuroscience

RIKEN Brain Science Institute

Computational Cognitive Neuroscience Unit (adjunct)

Kyoto University

RIKEN:
The Institute of Physical and Chemical Research



http://www.brain.riken.jp

President:
Hiroshi Matsumoto

Director:
Susumu Tonegawa

Nakahara Lab.

Job opportunities at BSI: http://www.brain.riken.jp/en/careers/

BSI Summer Program: http://www.brain.riken.jp/en/summer/

Foreign Postdoctoral Researcher Program: http://www.riken.jp/en/careers/programs/fpr/

International Program Associate: http://www.riken.jp/en/careers/programs/ipa/

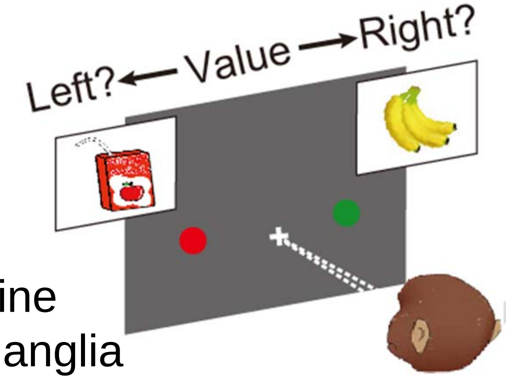
Our laboratory's interest

Computational principles linking brain mechanisms and behavior

Social decision-making

Decision-making

Reinforcement learning



Dopamine
Basal ganglia

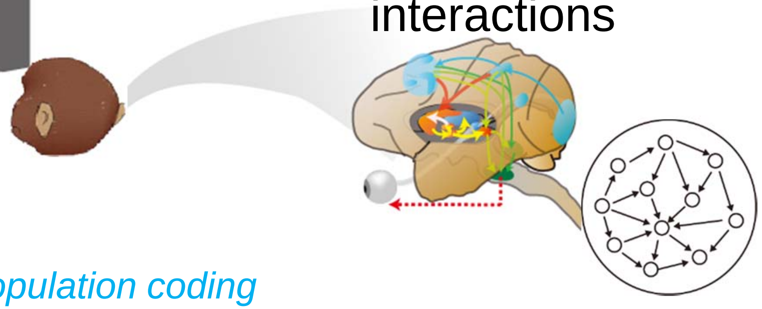
behavioral complexity

Neural Computation

neural complexity

Neural coding

Circuits & interactions



Population coding
Information geometry for interactions
Itemset mining; matrix balancing

Our efforts to expand the field

By way of demonstrations
“sequential, contextual (successive), and social”

Reminders: reinforcement learning, dopamine, basal ganglia

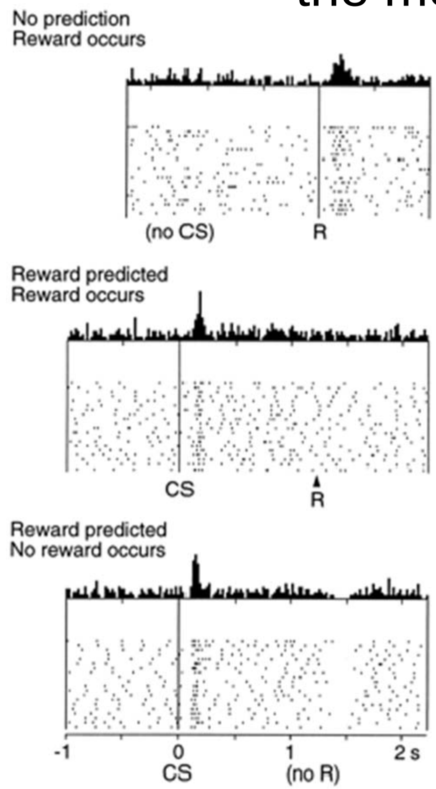
Dopamine (DA) activity

Reward prediction error hypothesis (Schultz et al. 1997)

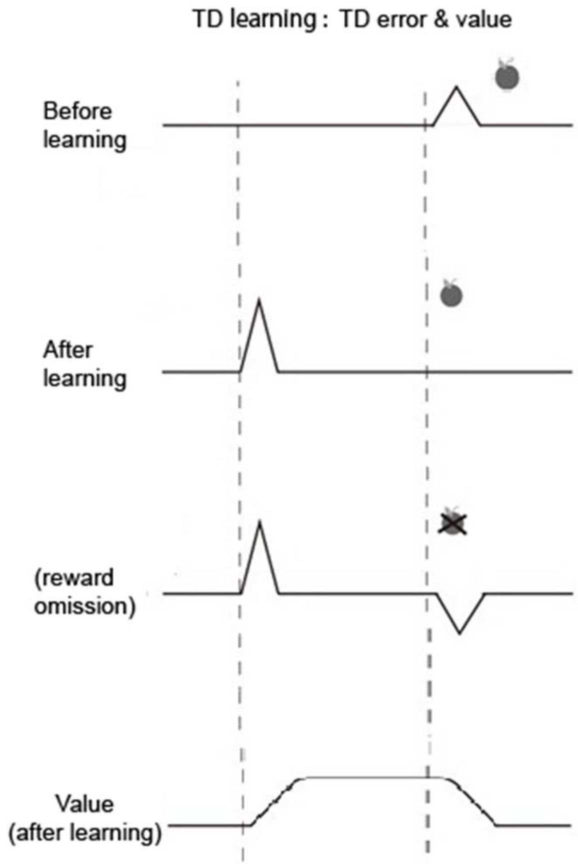
- *Signaling reward prediction error (= TD error)*
- *Functioning as TD learning signal*

Temporal difference learning (TD learning)

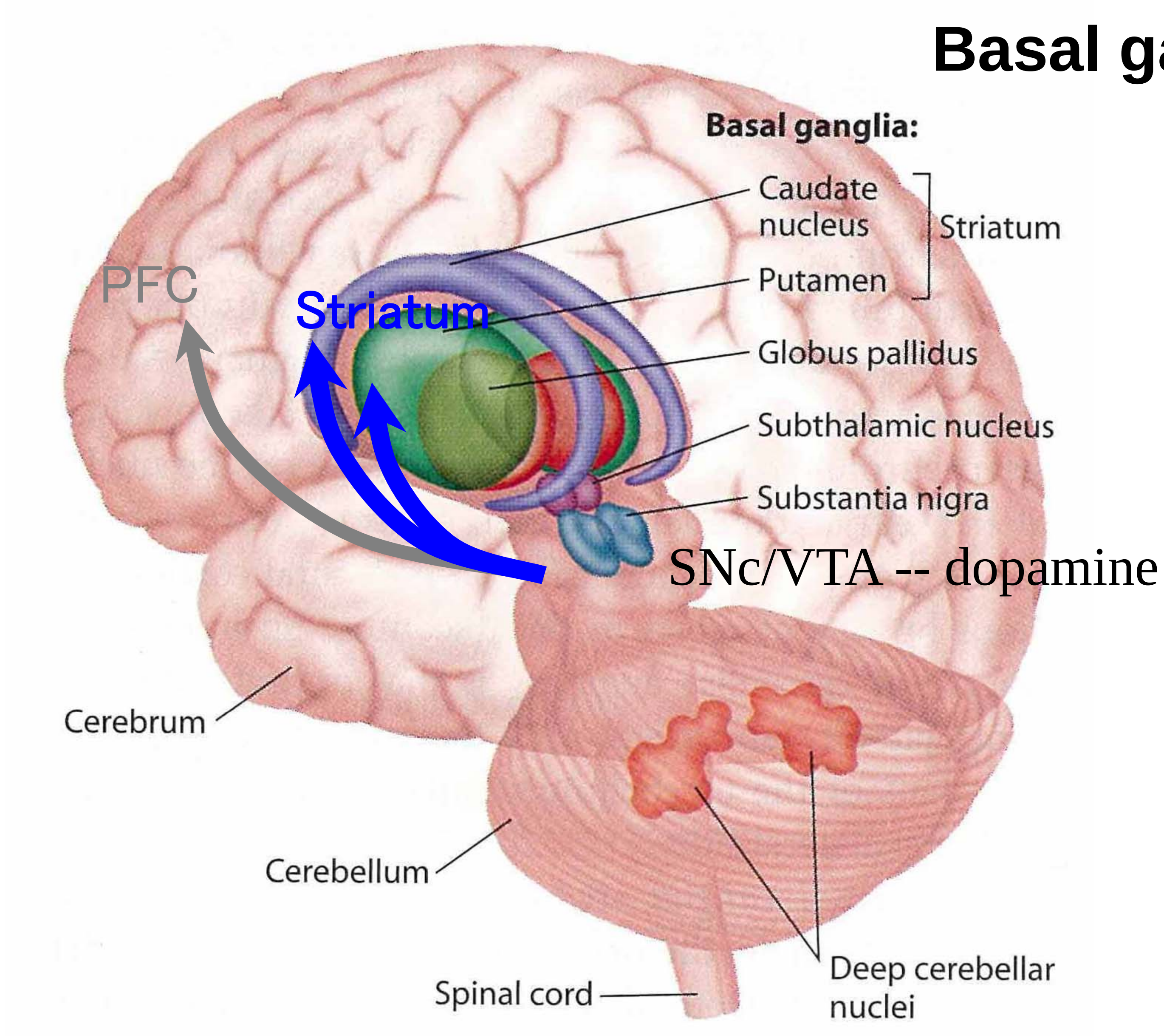
~ the most representative RL algorithm



W. Schultz (1998)



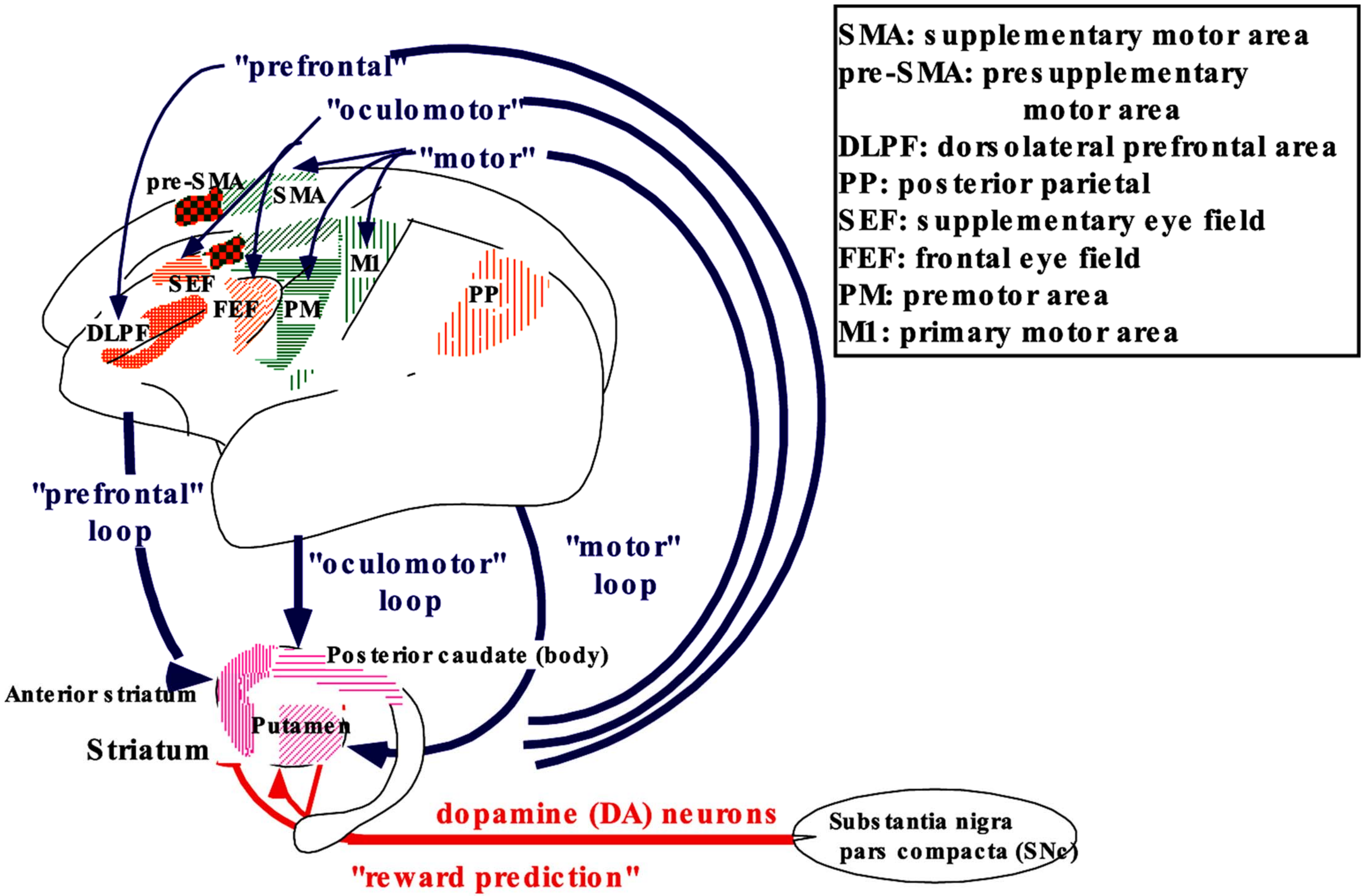
Basal ganglia



(Gazzaniga, cogniitve neuroscience)

Sequential

https://www.flickr.com/photos/vardolath/7321553930/sizes/l/



SMA: supplementary motor area
pre-SMA: presupplementary motor area
DLPF: dorsolateral prefrontal area
PP: posterior parietal
SEF: supplementary eye field
FEF: frontal eye field
PM: premotor area
M1: primary motor area
"prefrontal"
"oculomotor"
"motor"
pre-SMA
SMA
M1
SEF
FEF
PM
PP
DLPF
"prefrontal" loop
"oculomotor" loop
"motor" loop
Posterior caudate (body)
Anterior striatum
Putamen
Striatum
dopamine (DA) neurons
Substantia nigra pars compacta (SNc)
"reward prediction"

Experiment: 2x5 task

Serial Button Press task (Hikosaka et al., 95)

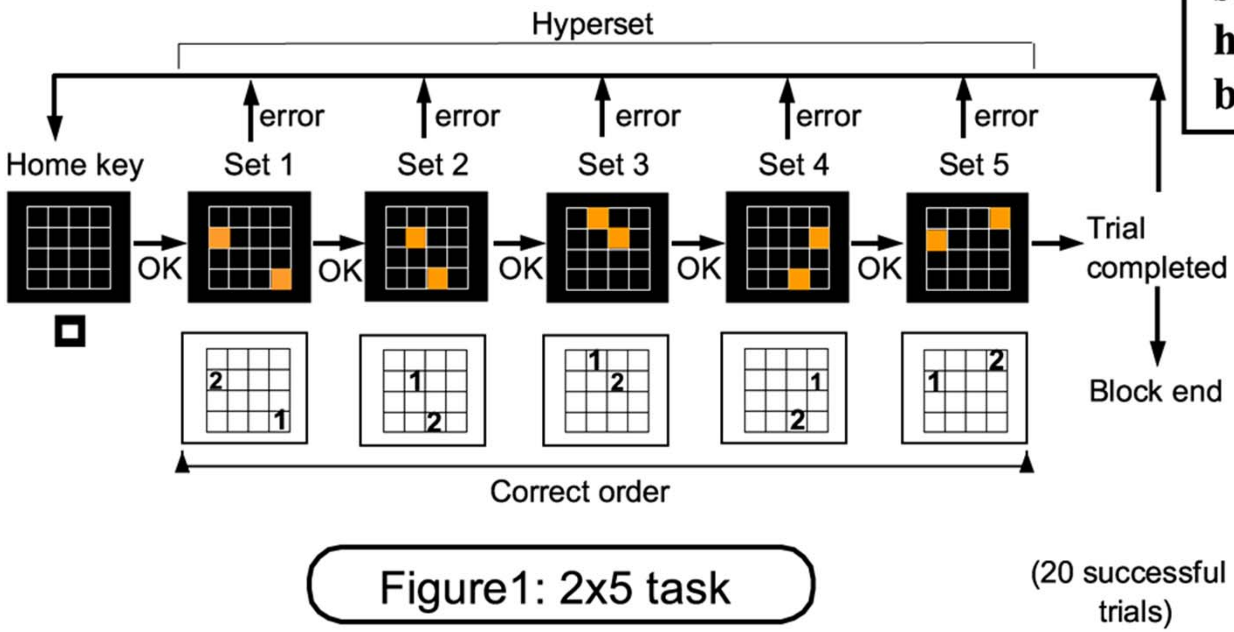


Figure1: 2x5 task

set: 2 buttons on 4x4 display
hyperset: 5 sets
block: 20 (or 10) successful trials

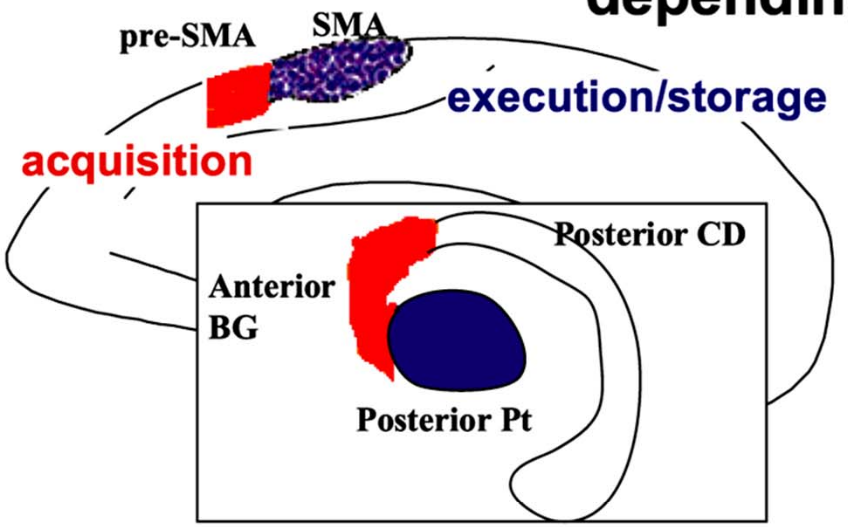
each day:
14 ~ 20 hypersets (blocks)

most of hypersets, experienced once (*new hypersets*) movie

some extensively experienced (*learned hypersets*)

Functional differentiation (blockade by muscimol injection)

The decrease of the performance varied depending on the blocked area.

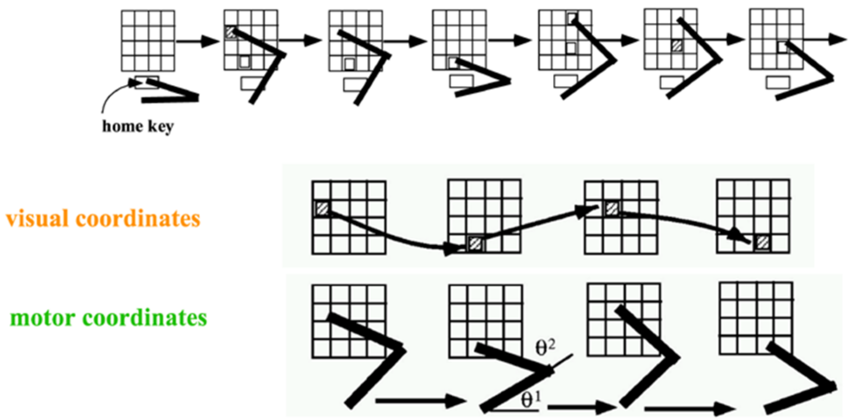


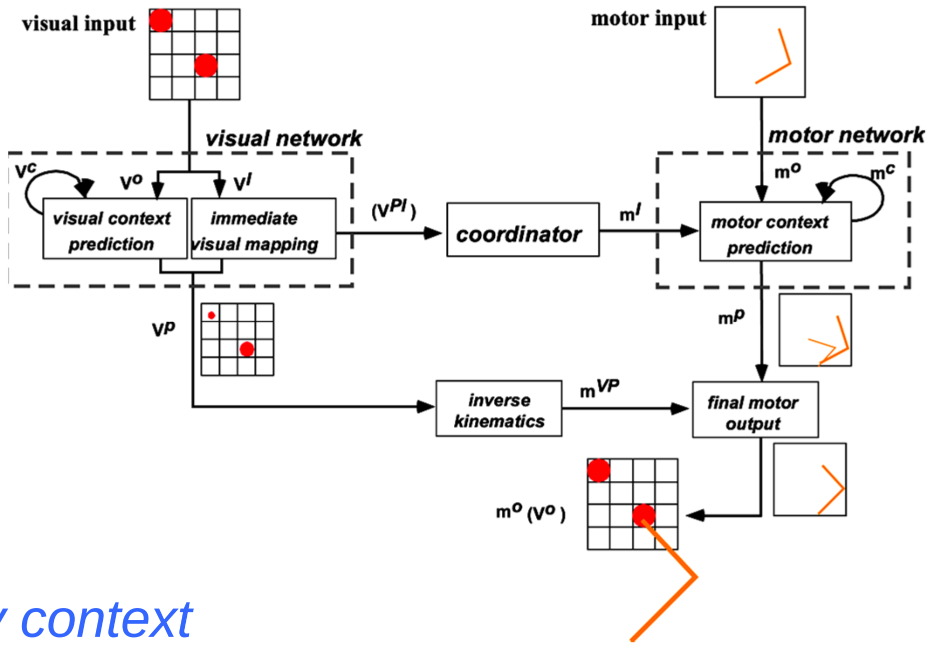
	Learned	New
anterior St	↓	↓ ↓
posterior St (pt)	↓ ↓	—
pre-SMA	—	↓ ↓ ↓
SMA	—	↓
	↓ (bilateral)	

The decrease of the perfor

Cortico-basal ganglia loops
-- parallel representation for sequence learning

Concurrent learning in visual and motor presentations
-- different computational advantage

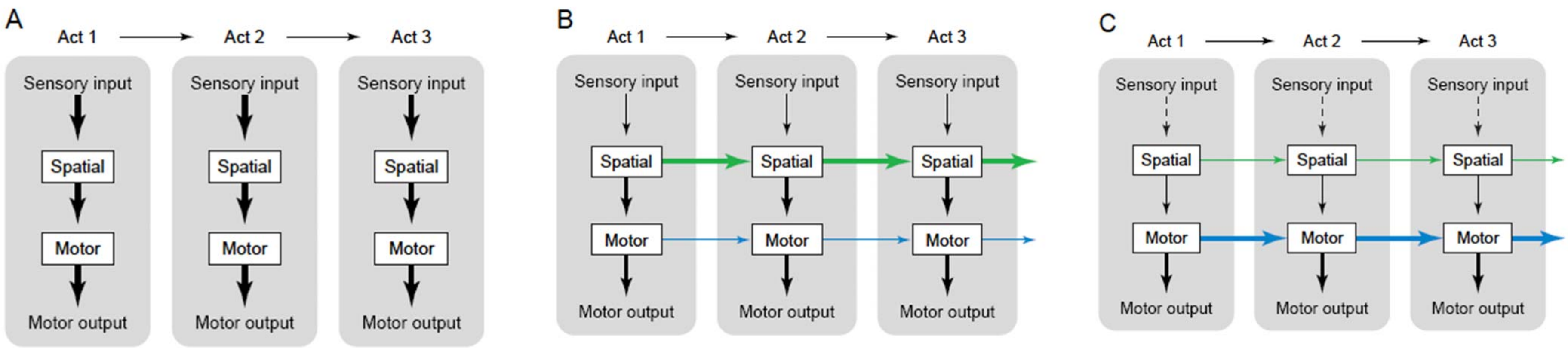




Prediction learning by context
Resetting of (ordinary) input-output actor

(Nakahara et al 2001)

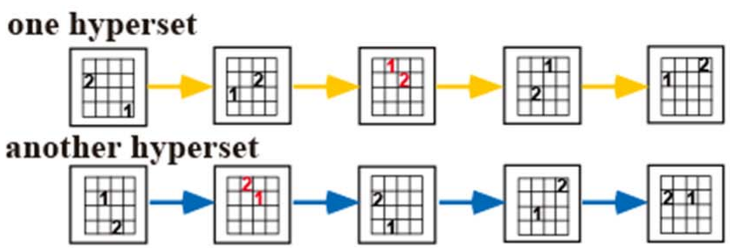
Three stages for sequence learning

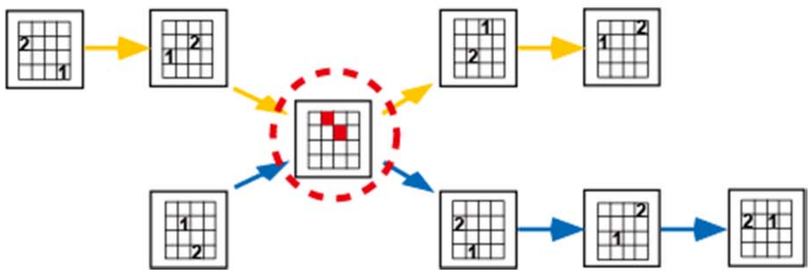


(Hikosaka, Nakahara et al 1999)

- Continuous control
- Similarity 3 stages – alpha GO
- Multiple systems in consort

Needs of context





(Nakahara et al 2001)

Results: validation by simulations

--- Learning process in training period

--- Role of each function module

--- Effect of resetting the immediate visual mapping

--- Reverse hyperset simulation

--- Opposite hand simulation

--- DA dysfunction simulation

--- Blockade simulations:

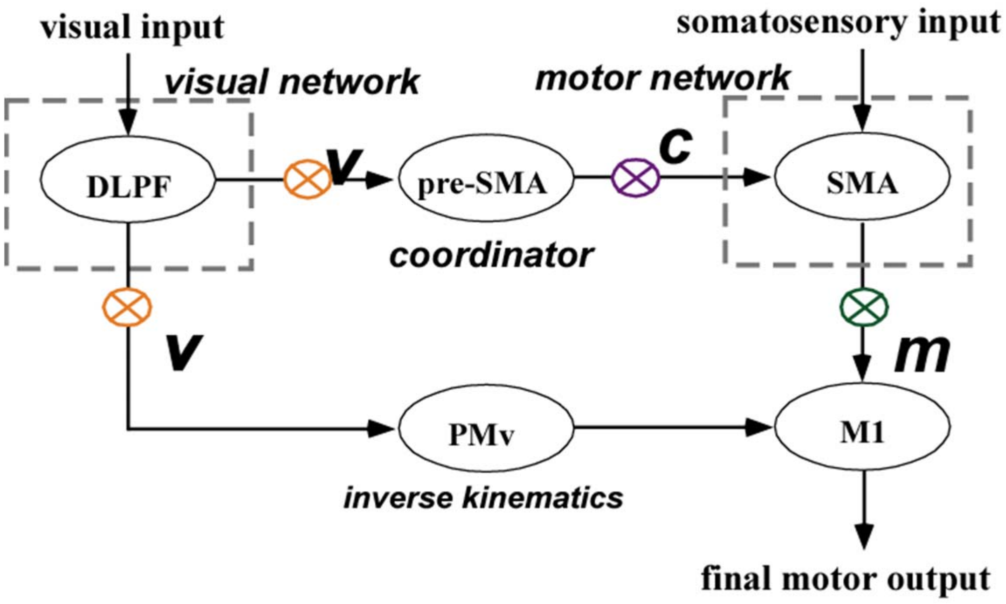
(1) the visual network, (2) the motor network, (3) the coordinator

Important to examine multiple aspects

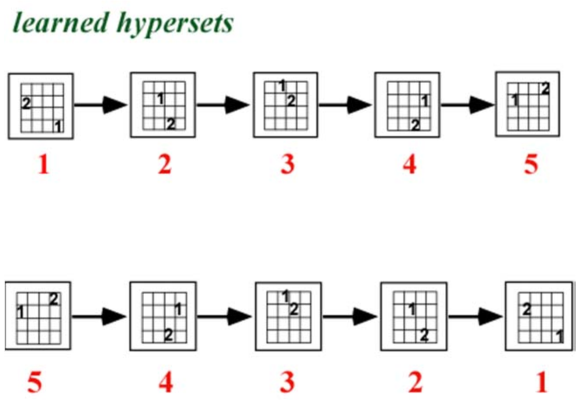
Only show a few of the results

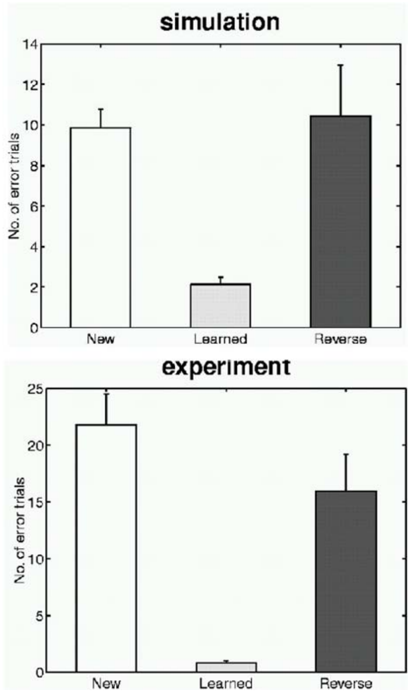
Blockade simulations

	performance	
blockade	Learned	New
visual loop	↓↓↓	↓↓↓↓
motor loop	↓↓↓↓	↓↓
coordinator	—	↓↓

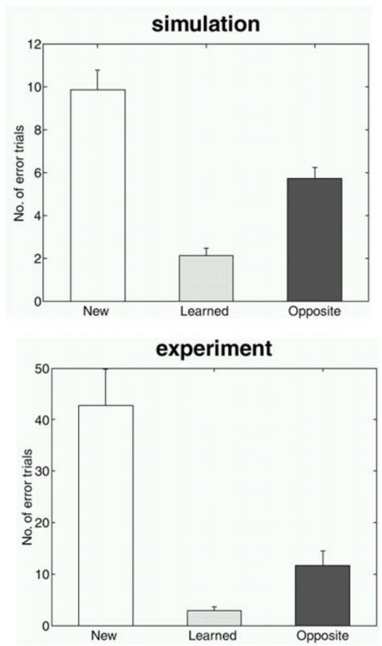


Reverse hyperset simulation





Opposite hand simulation

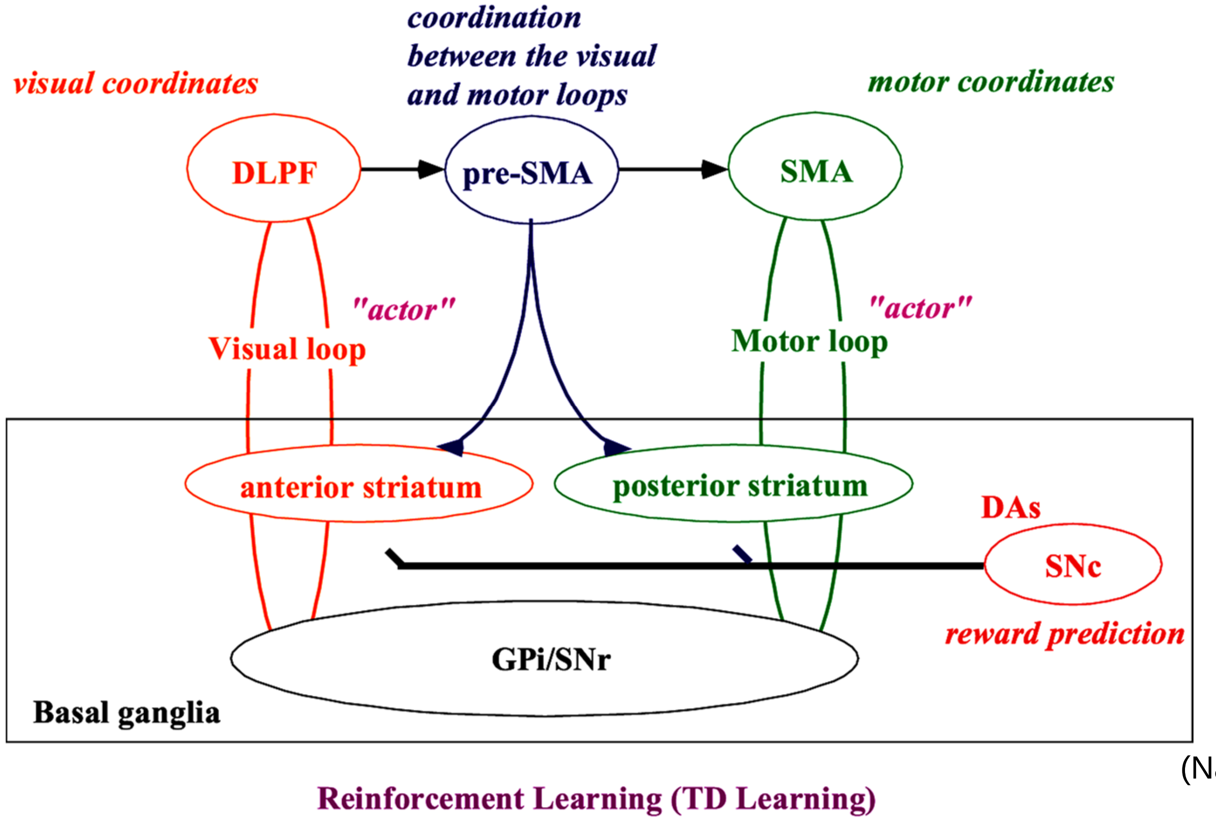


(Nakahara et al 2001)

Conclusion

Parallel representation hypothesis for sequential control and learning

Rapid acquisition & robust execution of sequences is realized by cooperation of the parallel BG loops, using different characteristics of different representations and learning signals of DA neurons.



(Nakahara et al 2001)

Contextual (successive)

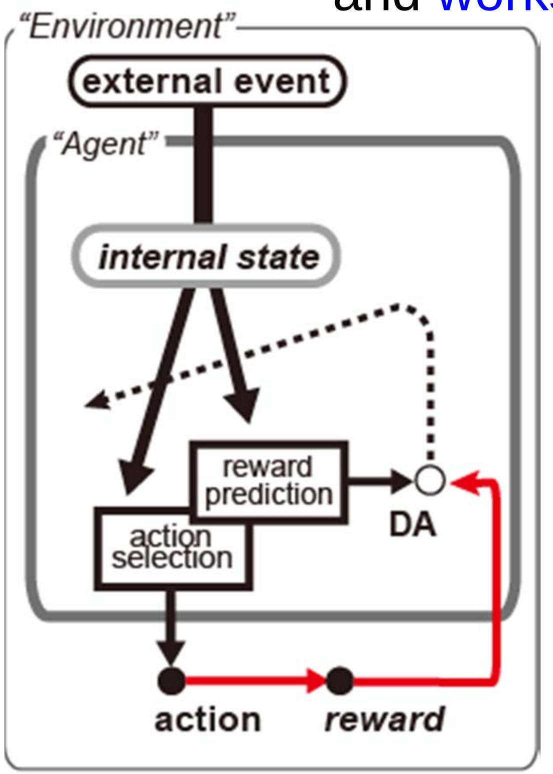
Where are we with our current understanding of neural RL, especially about dopamine?

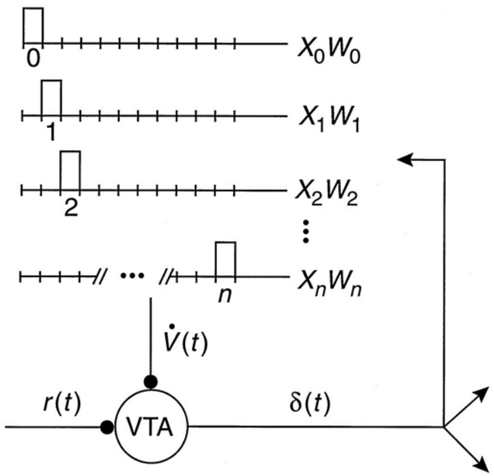
Our guiding hypothesis!!

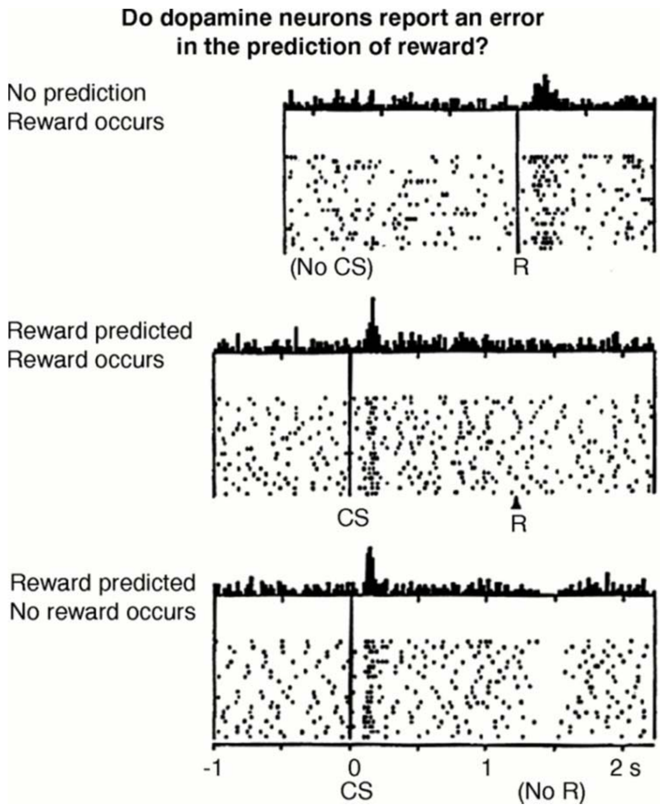
Reward prediction error hypothesis

Dopamine activity reports reward prediction error
and works as learning signal for reward prediction

(and action selection)

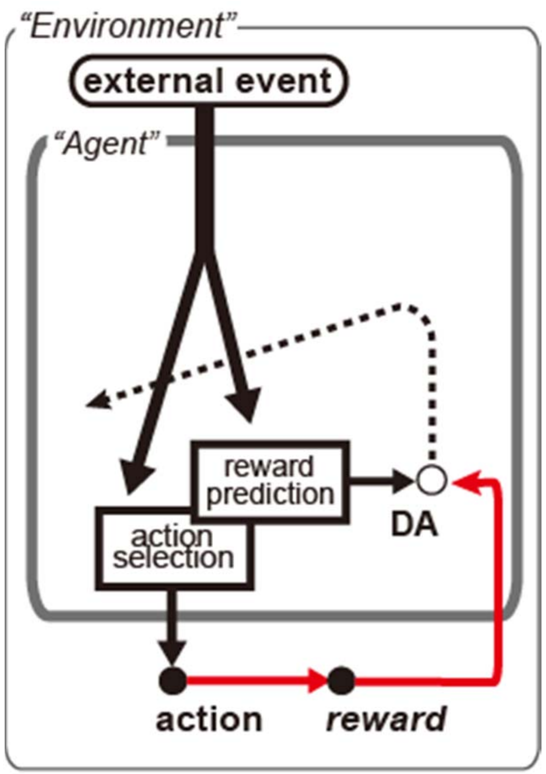


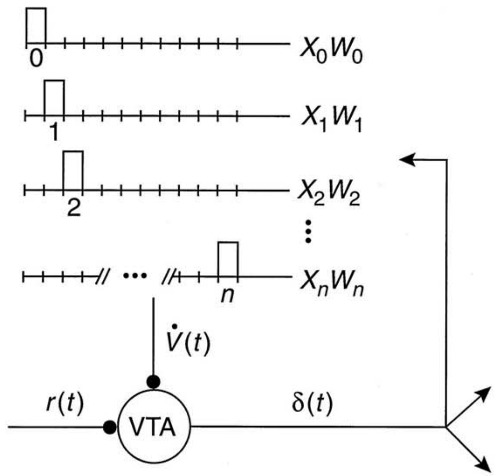


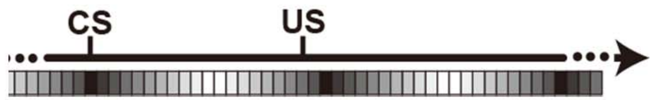


(Schultz et al 1997)

Reward prediction error hypothesis







In practice, for most of studies, information in each state is bounded by most recent sensory event.

In that sense, reward prediction of the hypothesis (in practice) is <u>a core, *or specific* prediction</u>

What if DA can report “better” RPE…

-- representation vs prediction
-- model-free vs model-based

Task 1 (non-contextual task)

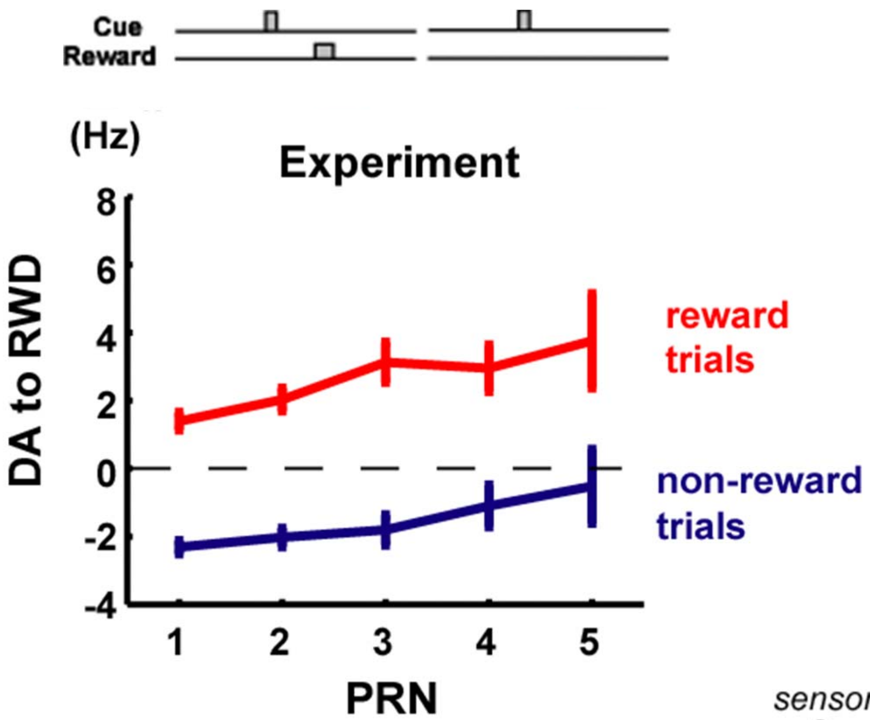
Classical conditioning task with 50 % reward probability

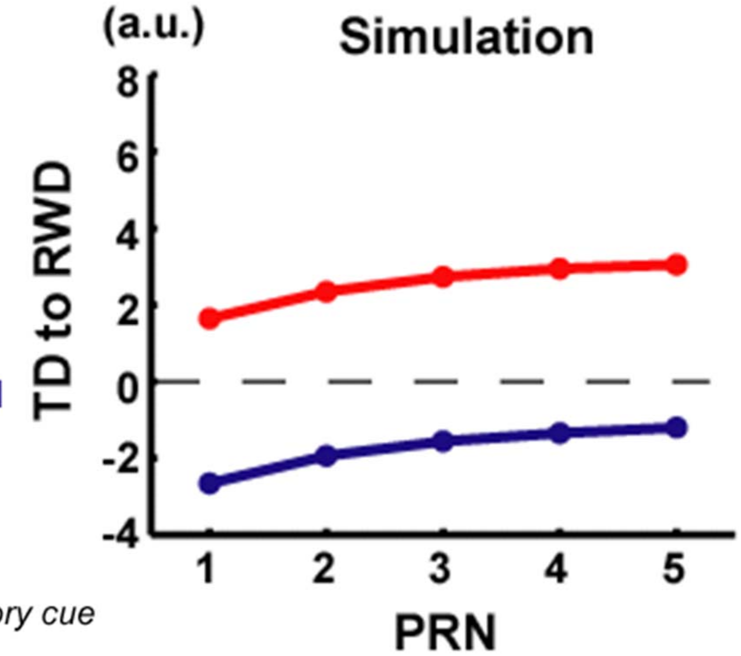
TD learning rule

$$\mathrm{TD} = r - V(s)$$

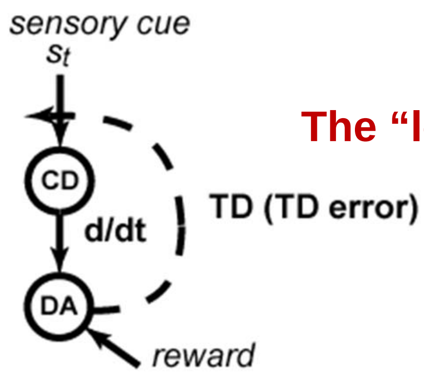
$$V(s) \rightarrow V(s) + \alpha\, \mathrm{TD}$$

DA response $\pm$50% reward prediction error





Why the positive slope?

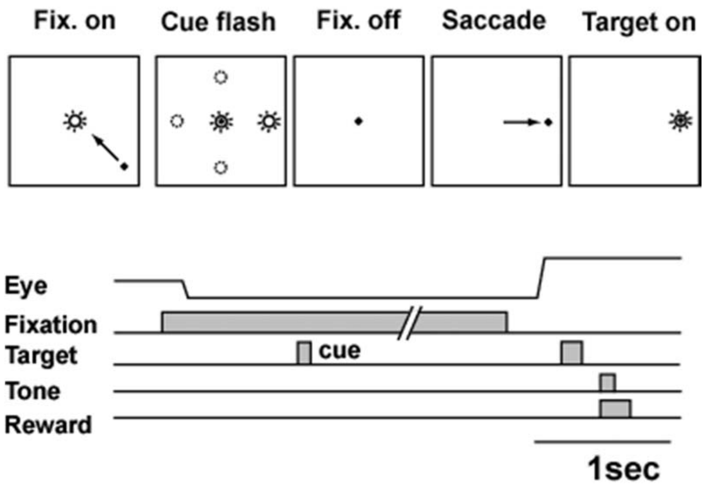


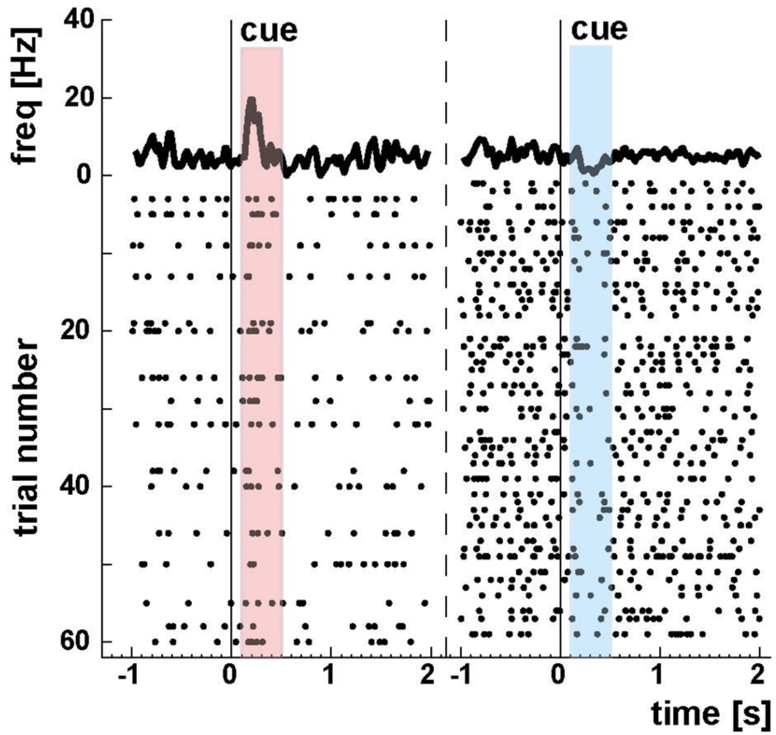
The “local” nature of TD learning

The slope should be *positive, given ‘default’ (conventional) TD*

Task 2 (contextual task)

Task: an asymmetrically-rewarded memory-guided saccade task ("1DR")



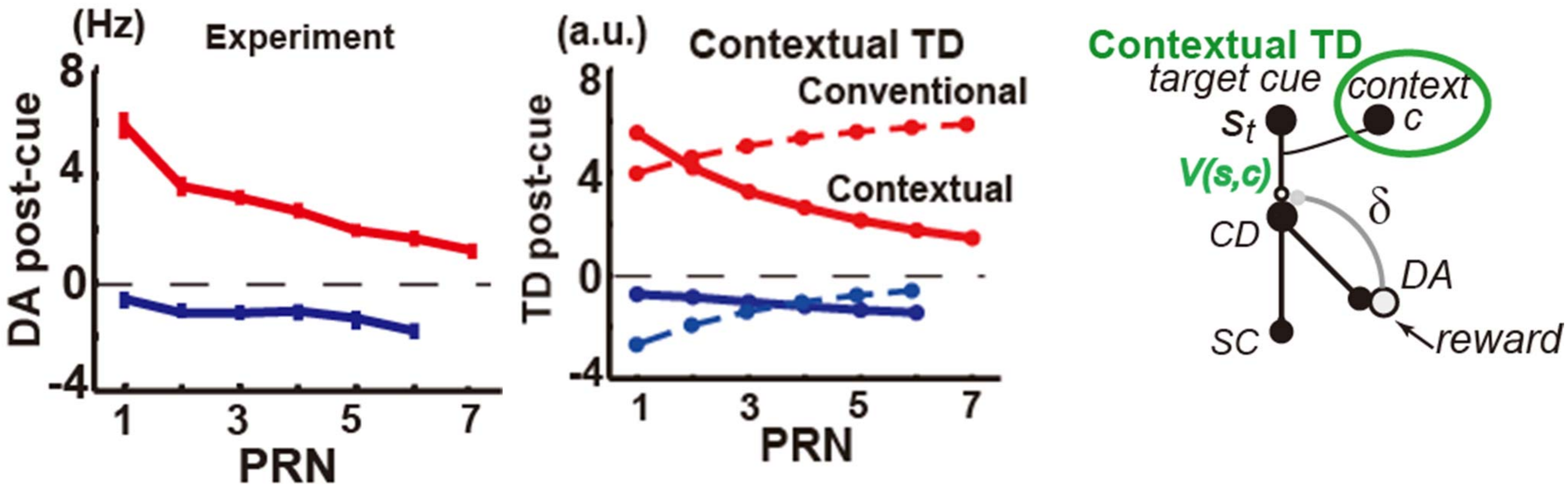


Ordinary probability, Pr (Rwd) = 1/4

However, hidden task structure exists

Sub-block	1				2				3			
Trial number	1	2	3	4	5	6	7	8	9	10	11	12
Target locations	L	R	U	D	U	L	D	R	D	U	L	R
Reward	-	+	-	-	-	-	-	+	-	-	-	+

DA activities can represent reward prediction error reflecting latent task structure over trials



Early instantiation: DA activity can be 'better' RPE error than the "default-model-free" RPE.

(in that sense, sort of "model-based" RPE w.r.t. the default) (Nakahara et al 2004, Neuron)

Representation and prediction of 'model-free'

Reward prediction error (RPE) as learning signal
⇔ Reward prediction (RP) learned

RP learned is limited by:
- information in RPE about reward statistics
- *state representation: capability to distinguish*

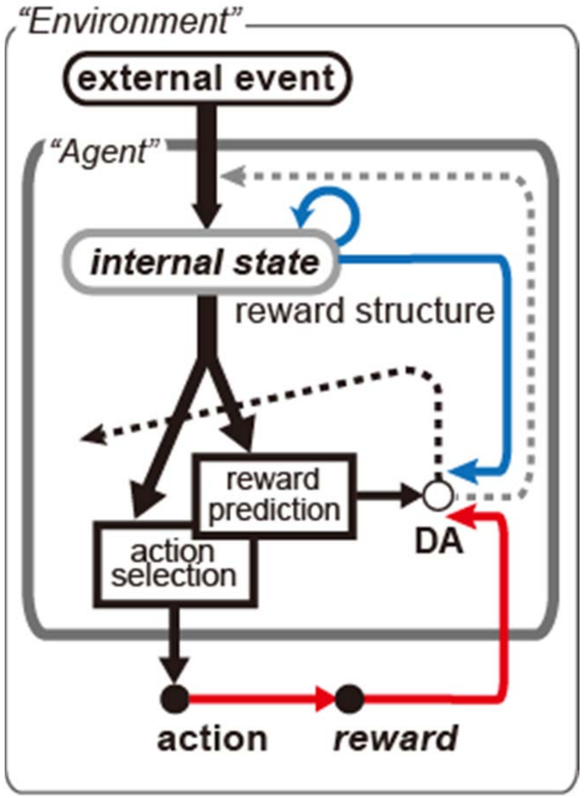
DA being better RPE than the default-model-free RPE

-- The better RPE is based on a better RP than the default RP.
-- The better RPE leads to a better model-free RP in learning

DA activity is a better TD error *Convention -- recent external event*
cf: LHb, Caudate
(Nakahara et al 2004; Bromberg-Martin et al 2010; Nakamura et al 2012)

Our suggestion:

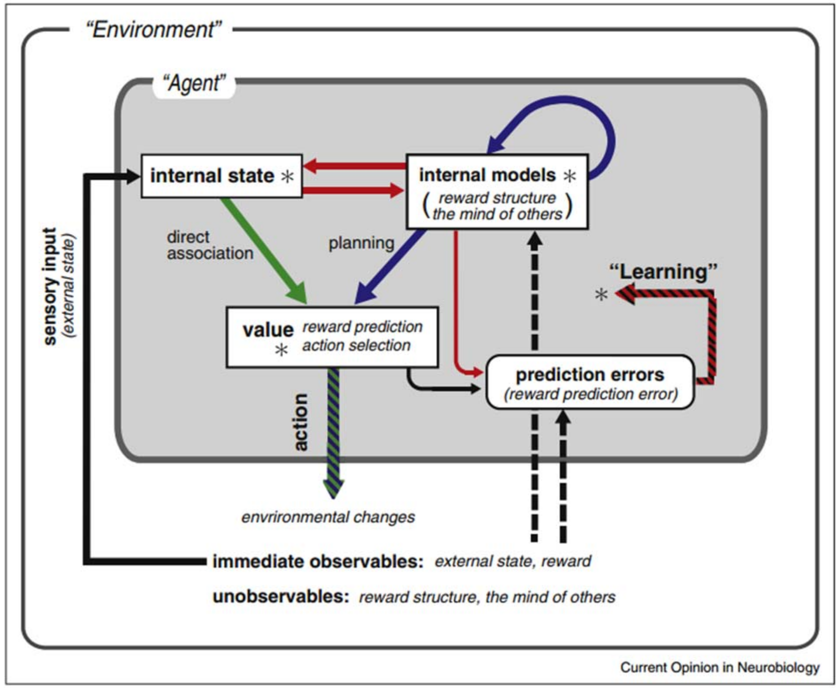
DA reward structural learning hypothesis



(Nakahara & Hikosaka, 2012)

Learning to represent reward structure, with better reward prediction

→ additional model-free / -based RL distinction required   cf) successor



(Nakahara 2014)

Social

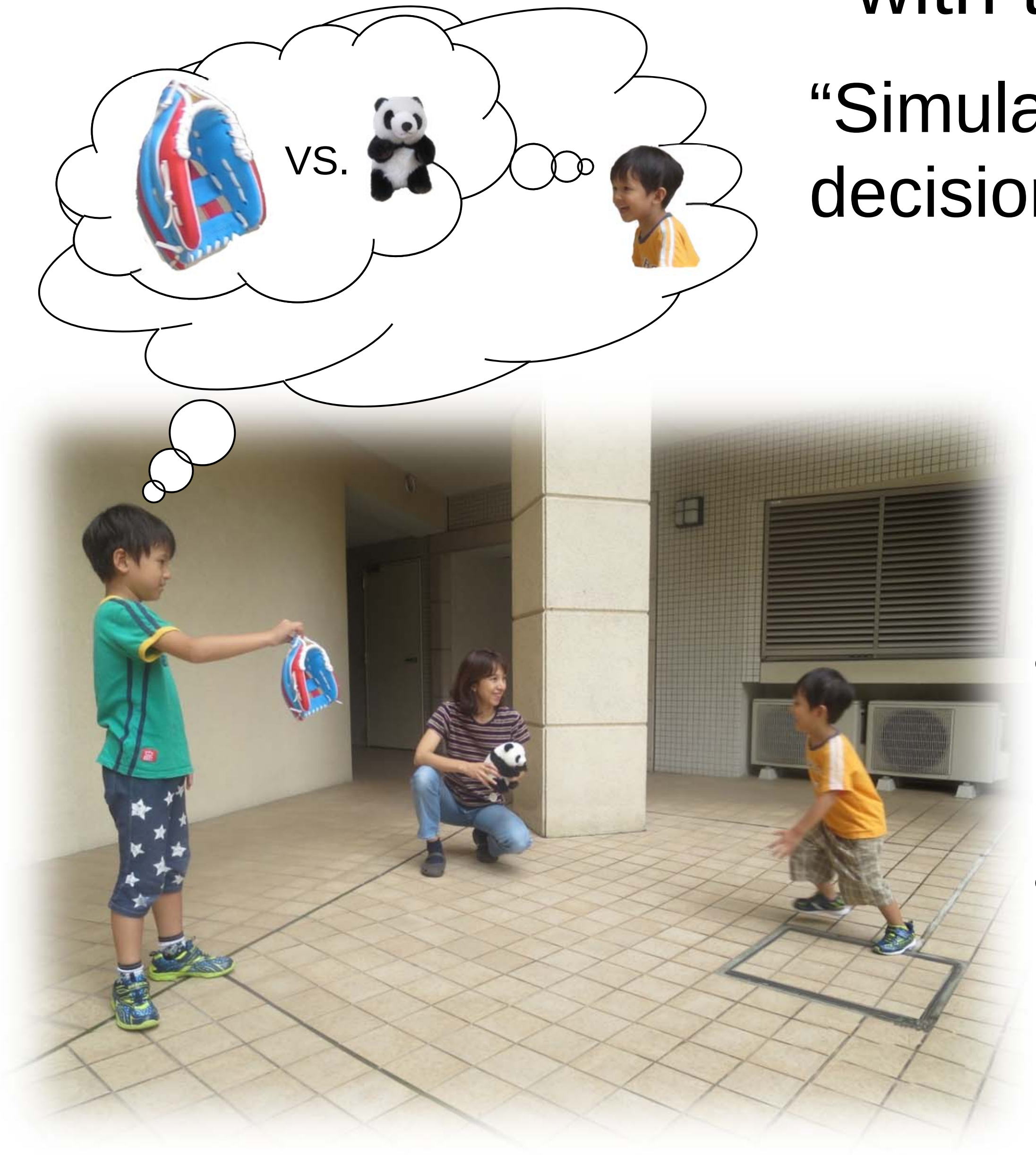
We like to understand brain functions, ultimately human brain functions.

Social behavior is a big part of being human

“Social”

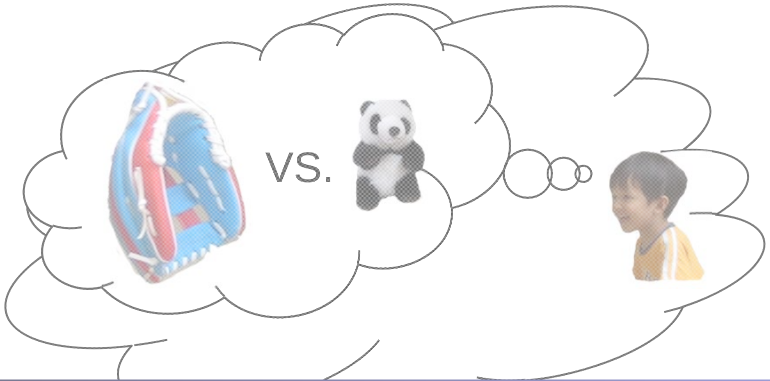
– with the **mind** of others

“Simulating” others’ internal decision-making process



- Apparently very complex
  - behavior and “mind”
- Computation is key

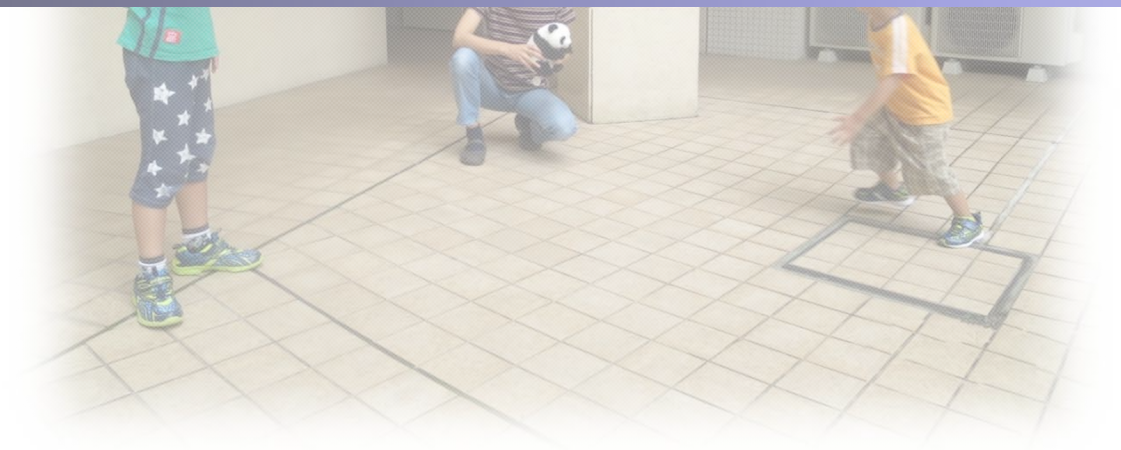
“Social”



– with the **mind** of others

“Simulating” others’ internal decision-making process

Can we develop quantitative understanding of social decision-making?

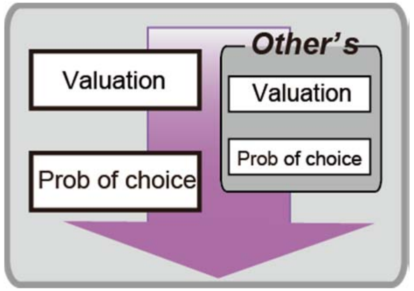


- Extend RL frameworks for quantitative social decision making
- Ask “social” questions *in* RL rather than apply RL to “social” Qs

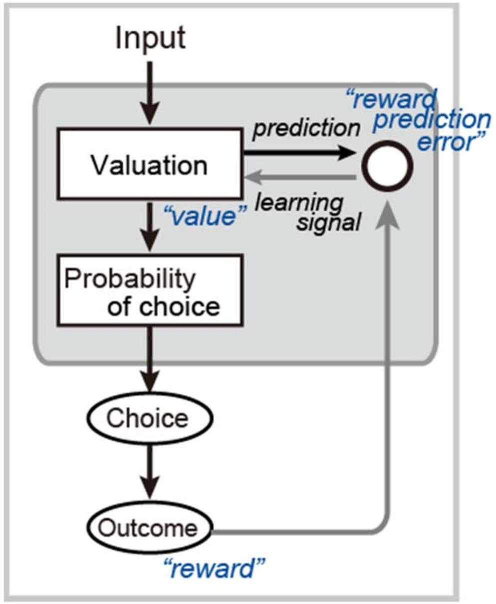
Key concepts in RL

Value (reward prediction) guides behavior.

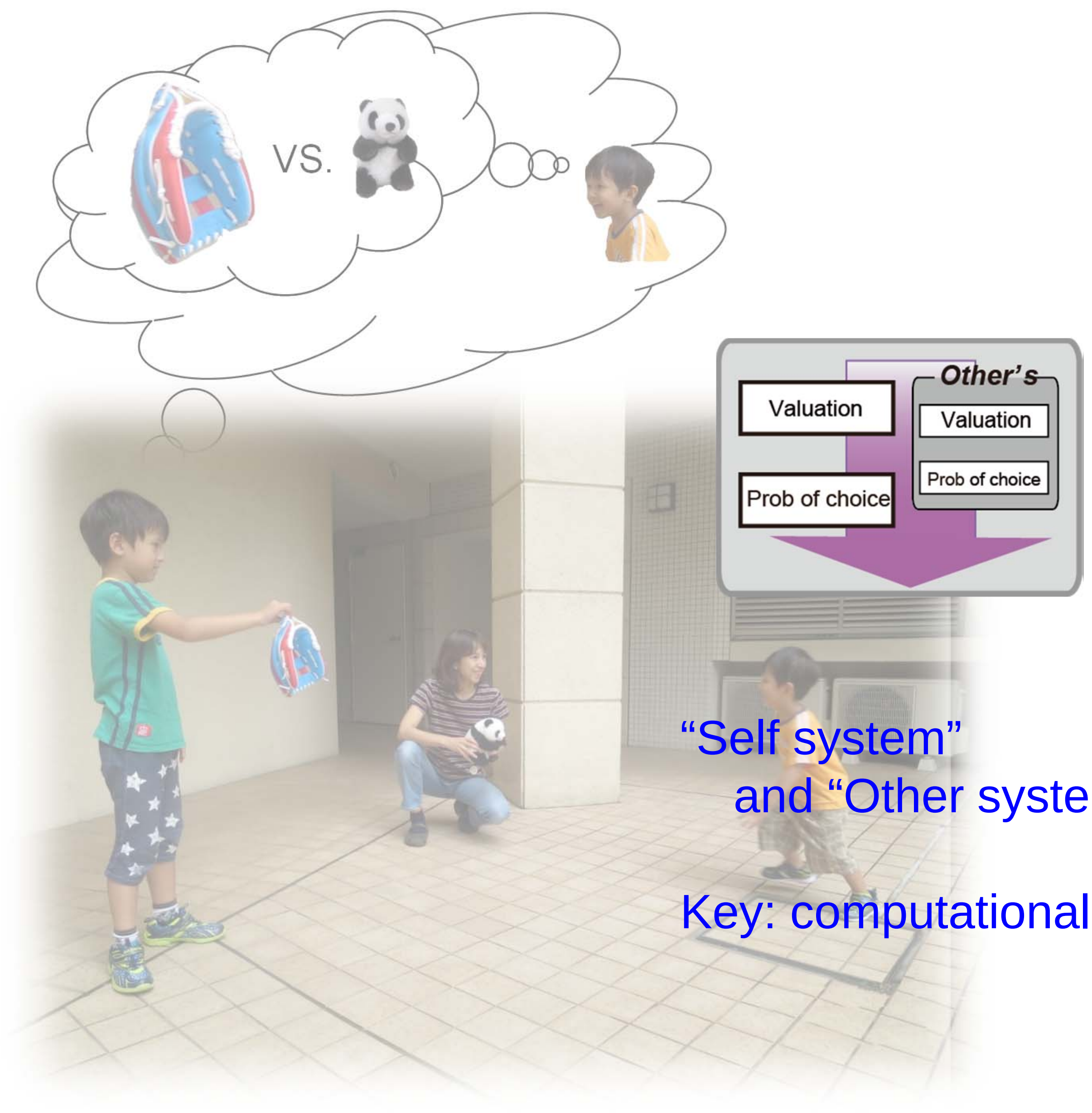
It is acquired using reward prediction error



“Self system” and “Other system”

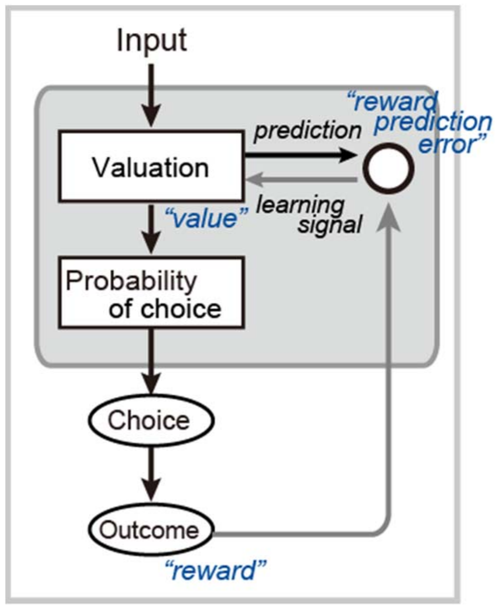


Key: computational primitives



“Self system”
and “Other system”

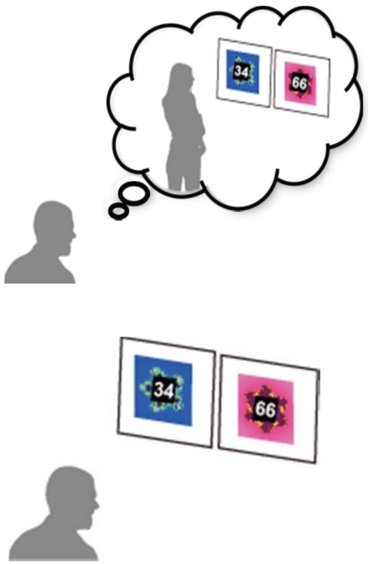
Key: computational primitives

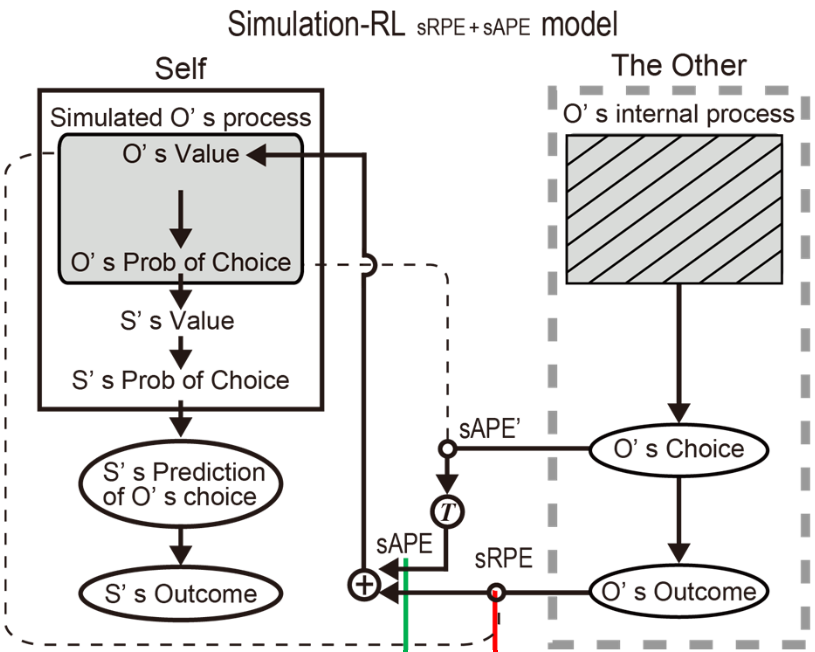


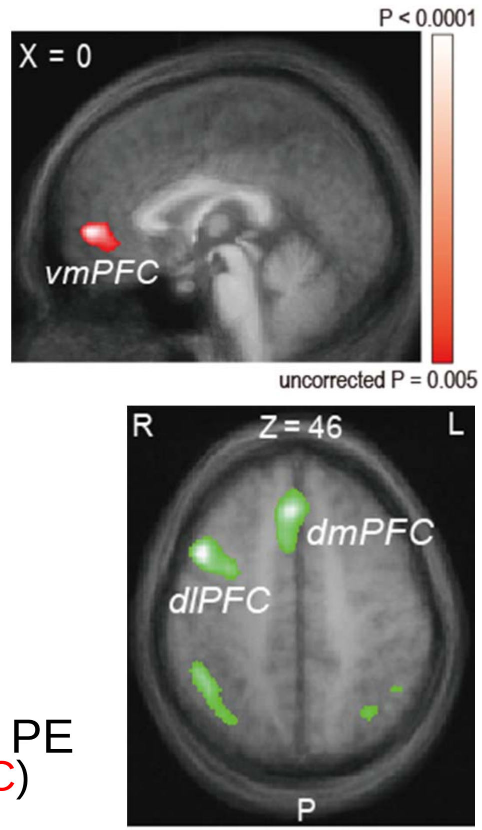
Learning to simulate others' decisions

Two simulation-learning signals:

"We are the same" and "We are different"







Simulated-other's Reward PE
(vmPFC)

Simulated-other's Action PE
(dmPFC/dlPFC)

S. Suzuki et al, *Neuron* (2012).

Thank you