OIST Computational Neuroscience Course 2024, June 22

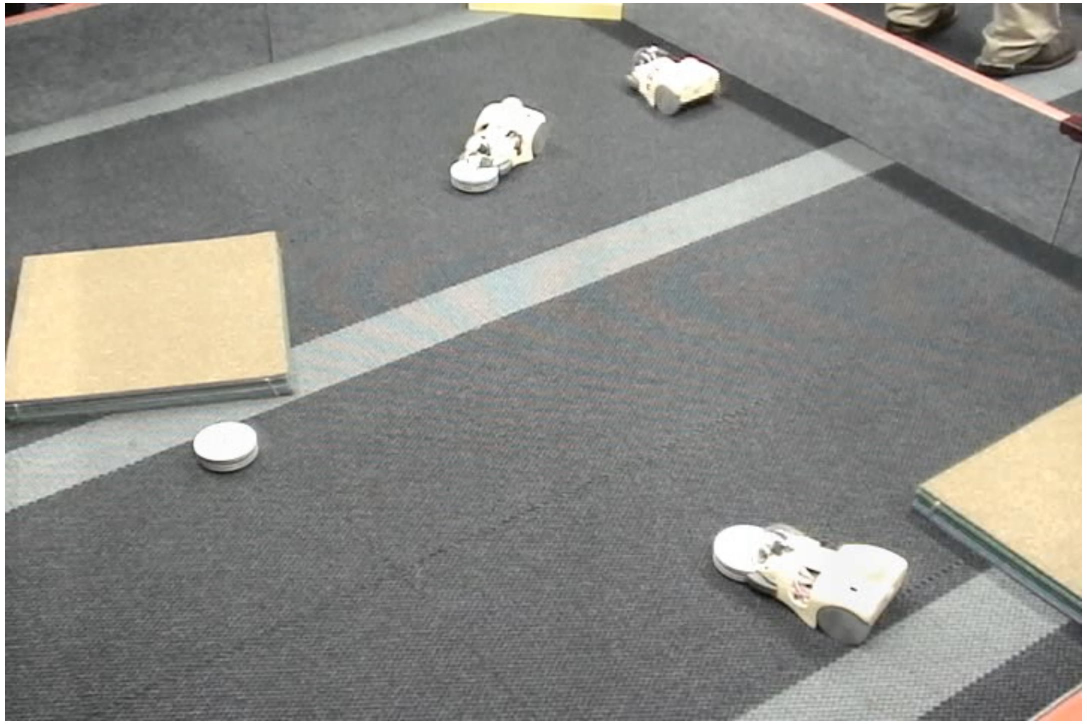
Reinforcement Learning and Bayesian Inference

Kenji Doya doya@oist.jp
Okinawa Institute of Science and Technology Graduate University
Neural Computation Unit
groups.oist.jp/ncu

OIST Neural Computation Unit

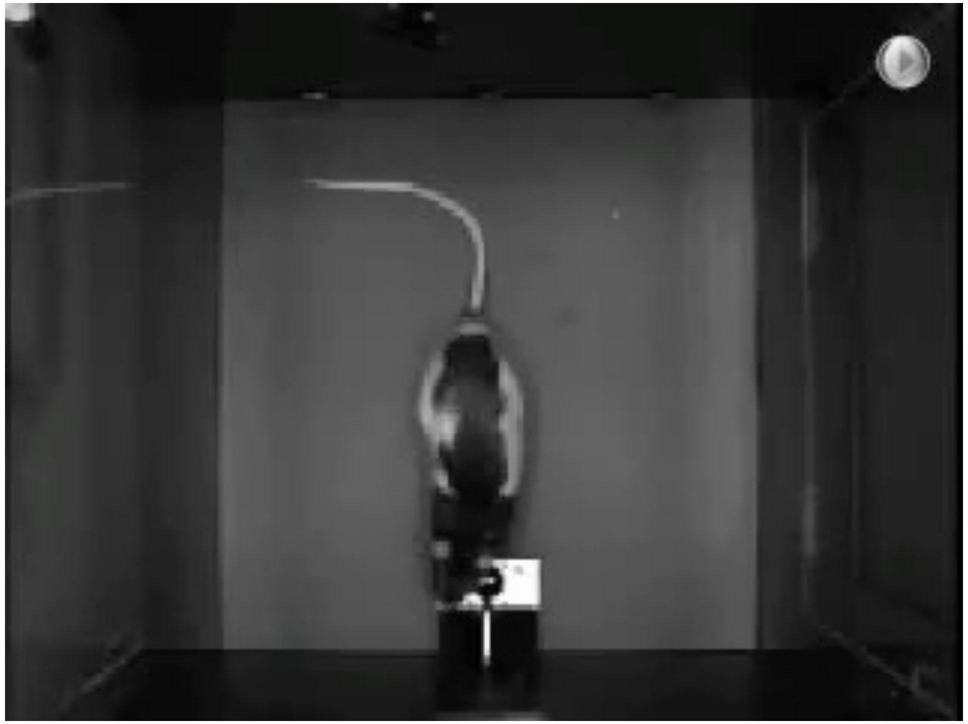
Create flexible learning systems

- robot experiments

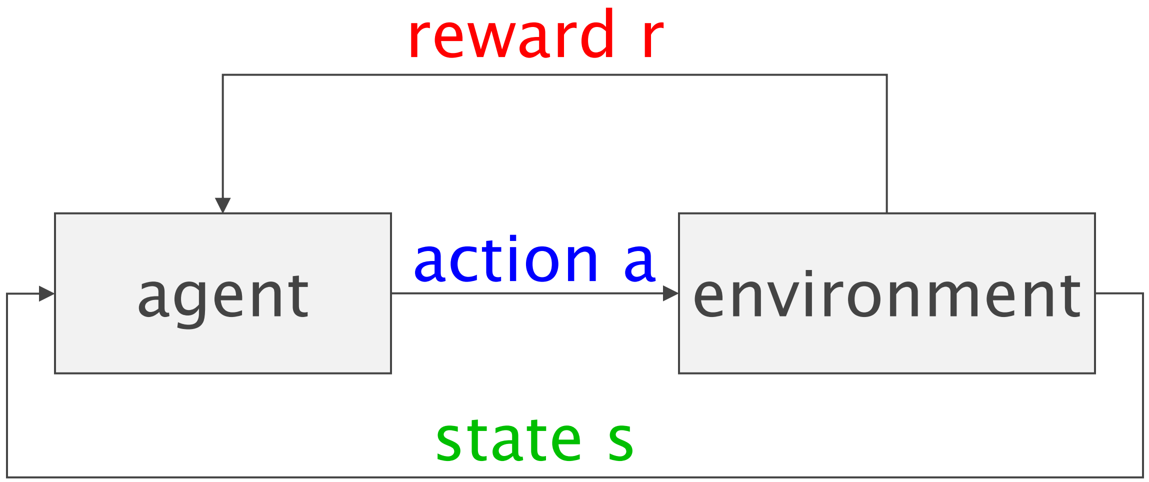


Reveal brain's learning mechanisms

- neurobiology



Reinforcement Learning



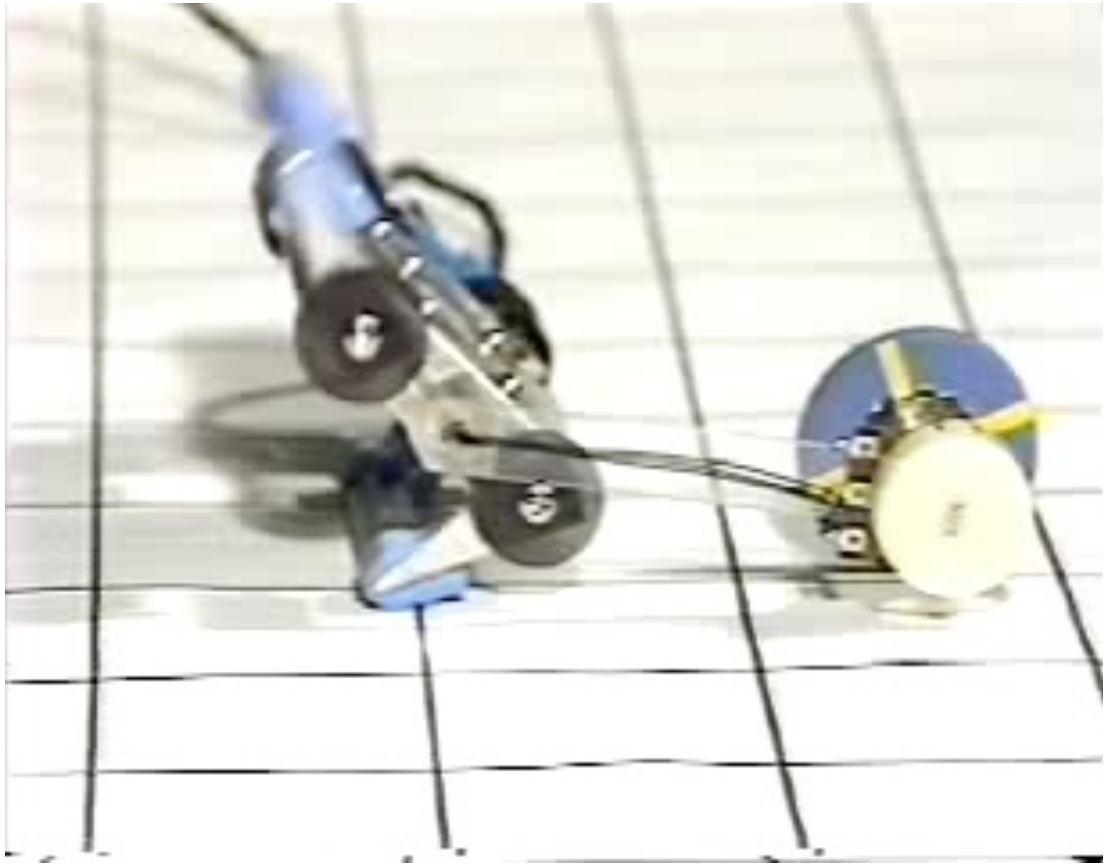
Learn action policy: s → a to maximize rewards

- Efficient algorithms for artificial agents
- Circuit and molecular mechanisms in the brain

Learning to Walk

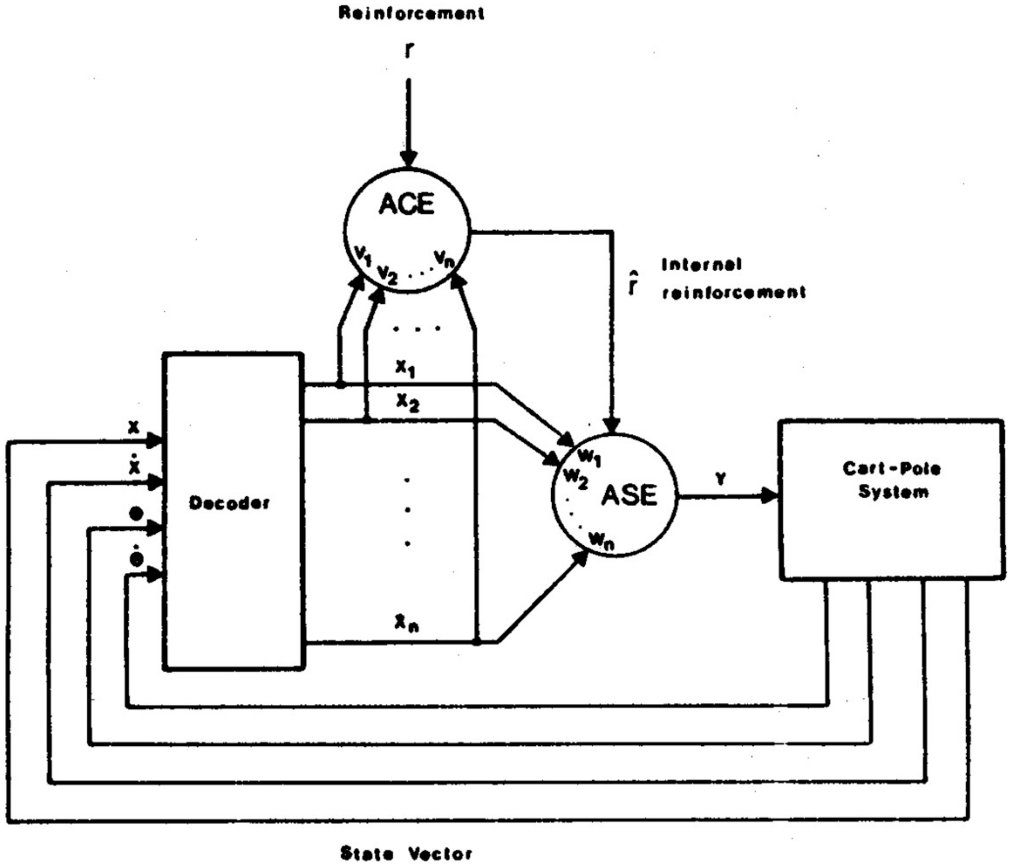
(Doya & Nakano, 1985)

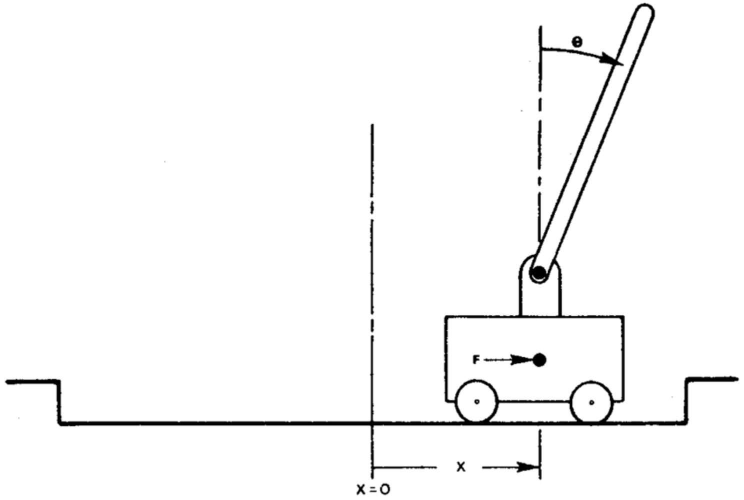
- Explore actions (cycle of 4 postures)
- Learn from performance feedback (speed sensor)

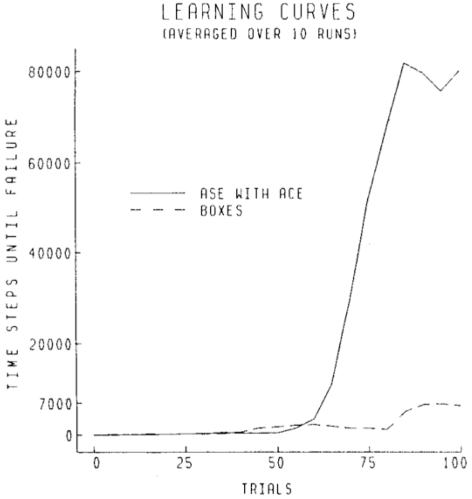


Neuronlike Adaptive Elements That Can Solve Difficult Learning Control Problems

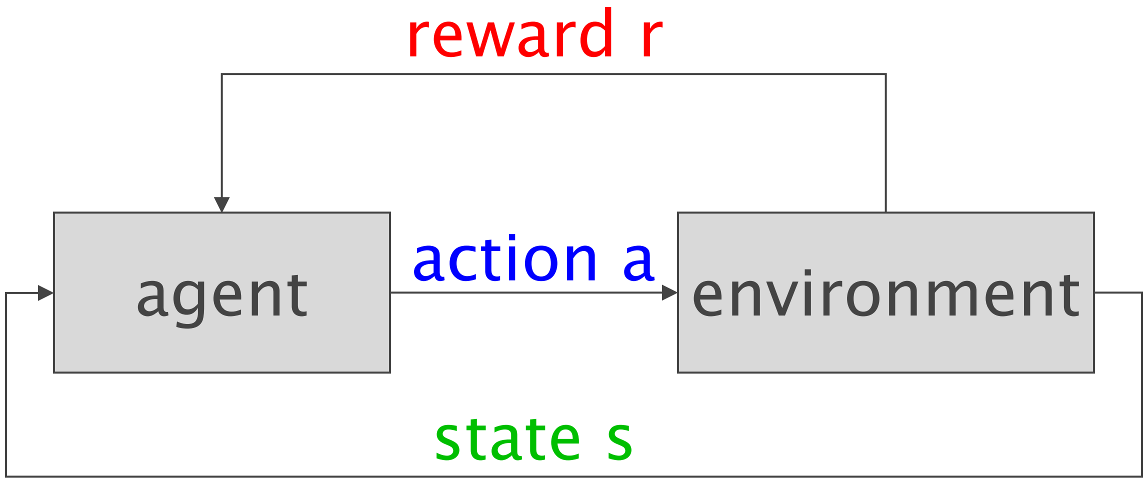
ANDREW G. BARTO, MEMBER, IEEE, RICHARD S. SUTTON, AND CHARLES W. ANDERSON (1983)







Markov Decision Process (MDP)



- Markov decision process
  - state $s \in \mathcal{S}$
  - action $a \in \mathcal{A}$
  - policy $p(a|s)$
  - reward $p(r|s,a)$
  - dynamics $p(s'|s,a)$
- Optimal policy: maximize cumulative reward
  - finite horizon: $E[\, r(1) + r(2) + r(3) + \ldots + r(T)]$
  - infinite horizon: $E[\, r(1) + \gamma r(2) + \gamma^2 r(3) + \ldots\,]$

    $0 \leq \gamma \leq 1$: temporal discount factor
  - average reward: $E[\, r(1) + r(2) + \ldots + r(T)]/T,\ T \rightarrow \infty$

Actor-Critic and TD learning

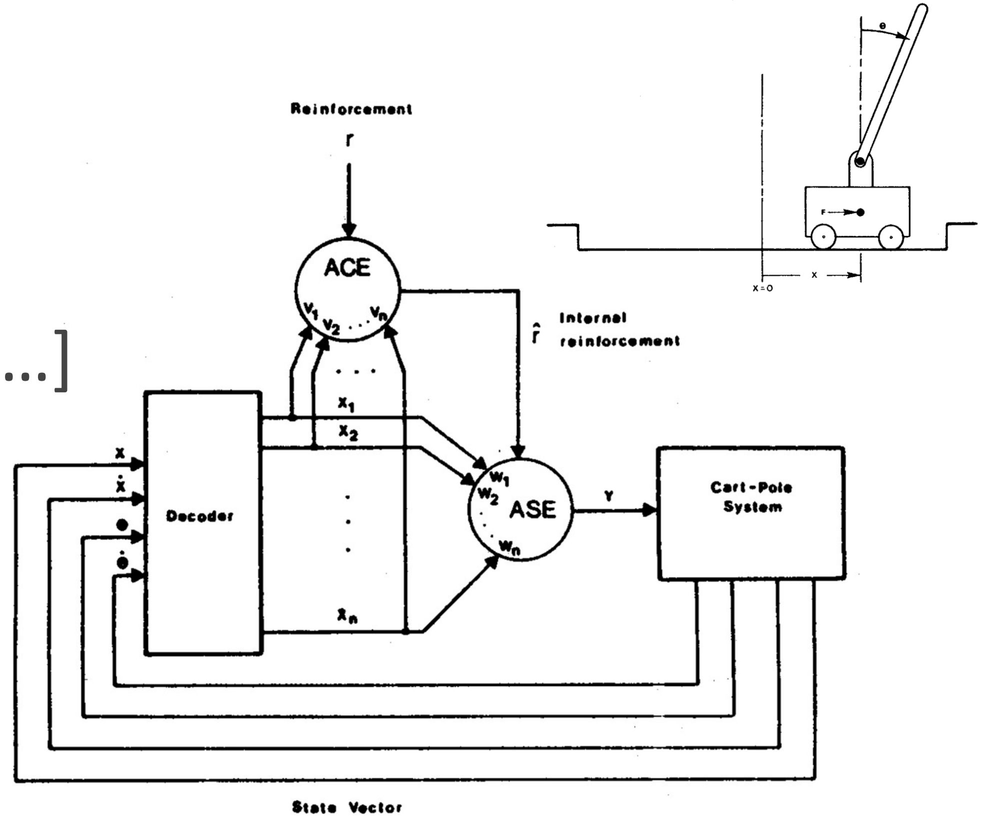
- Actor: policy with parameter w

  e.g., $a(t) = \Sigma_j w_j s_j(t) + \sigma n(t)$
- Critic: learn state value function
  - $V(s(t)) = E[\, r(t) + \gamma r(t+1) + \gamma^2 r(t+2) + \ldots]$

  e.g., $V(s(t);v) = \Sigma_j v_j s_j(t)$
- Temporal Difference (TD) error:
  - $\delta(t) = r(t) + \gamma V(s(t+1)) - V(s(t))$
- Critic learning: $\Delta V(s(t)) \propto \delta(t)$

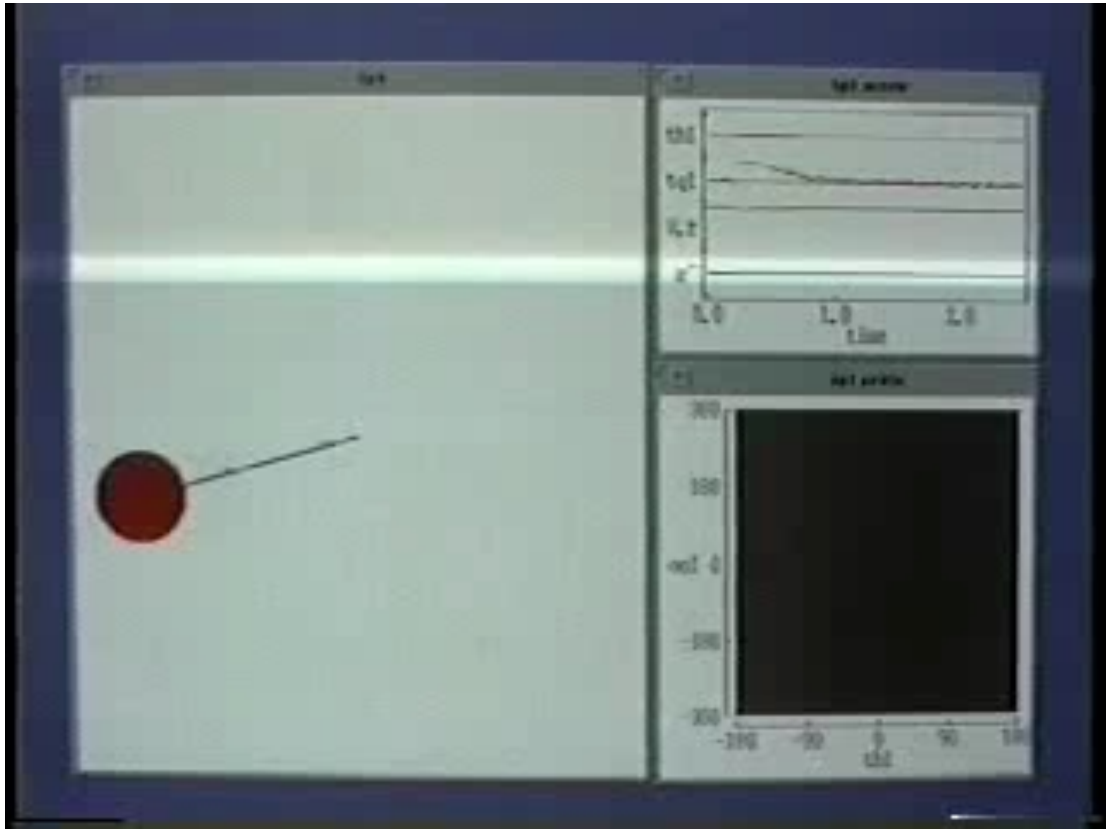
  $\Delta v_j = \alpha\, \delta(t)\, s_j(t)$
- Actor learning: $\Delta w \propto \delta(t)\, \partial \log P(a(t)|s(t);w)/\partial w$

  $\Delta w_j = \alpha_a\, \delta(t)\, \{a(t) - \Sigma_j w_j s_j(t)\}\, s_j(t)$ … weighted Hebb



Pendulum Swing-Up

- state: angle θ, angular velocity ω
- reward function: potential energy: cos θ



ω

θ

- Value function

SARSA and Q Learning

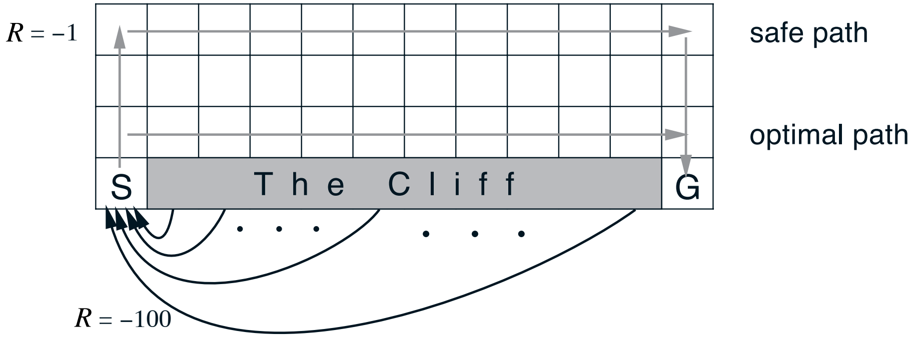
- Action value function
  - $Q(s,a) = E[\ r(t) + \gamma r(t+1) + \gamma^2 r(t+2) \ldots |\ s(t)=s, a(t)=a]$
- Action selection
  - ε-greedy: $a = \mathrm{argmax}_a\ Q(s,a)$ with prob $1-\varepsilon$
  - Boltzman: $P(a_i|s) = \exp[\beta Q(s,a_i)] / \Sigma_j \exp[\beta Q(s,a_j)]$
- Update by temporal difference (TD) error
  - $\Delta Q(s(t),a(t)) = \alpha\ \delta(t)$
  - SARSA: on-policy

    $\delta(t) = r(t) + \gamma Q(s(t+1),a(t+1)) - Q(s(t),a(t))$
  - Q learning: off-policy

    $\delta(t) = r(t) + \gamma \max_{a'} Q(s(t+1),a') - Q(s(t),a(t))$

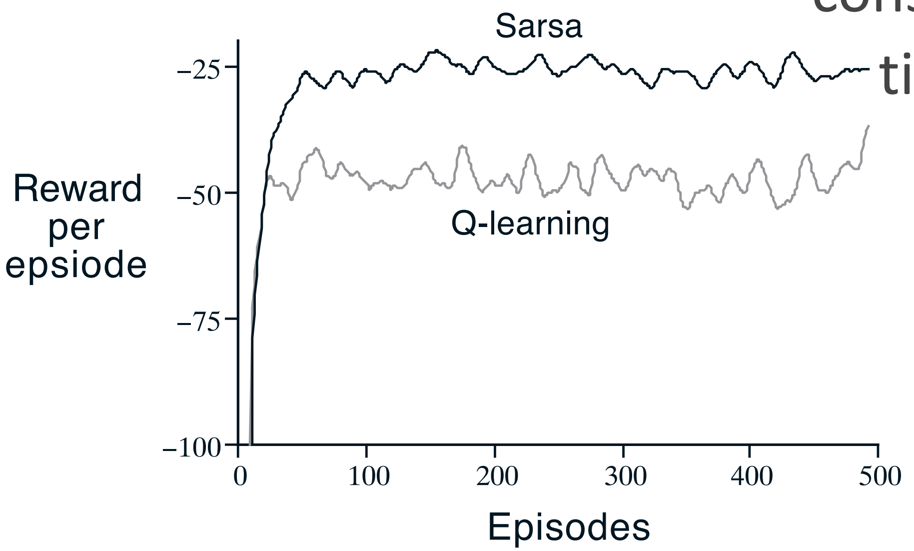
SARSA and Q Learning

- Cliff walking task (Sutton & Barto, 1998)



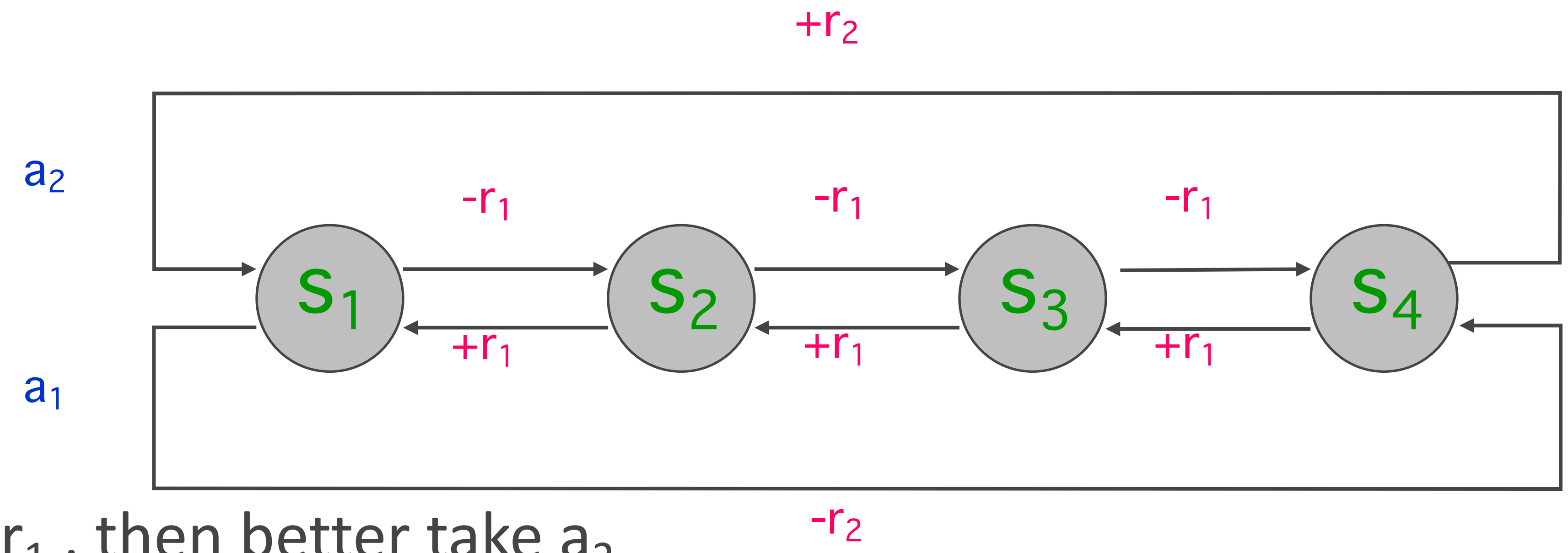
consider exploration

timal greedy policy



“Pain-Gain” Task

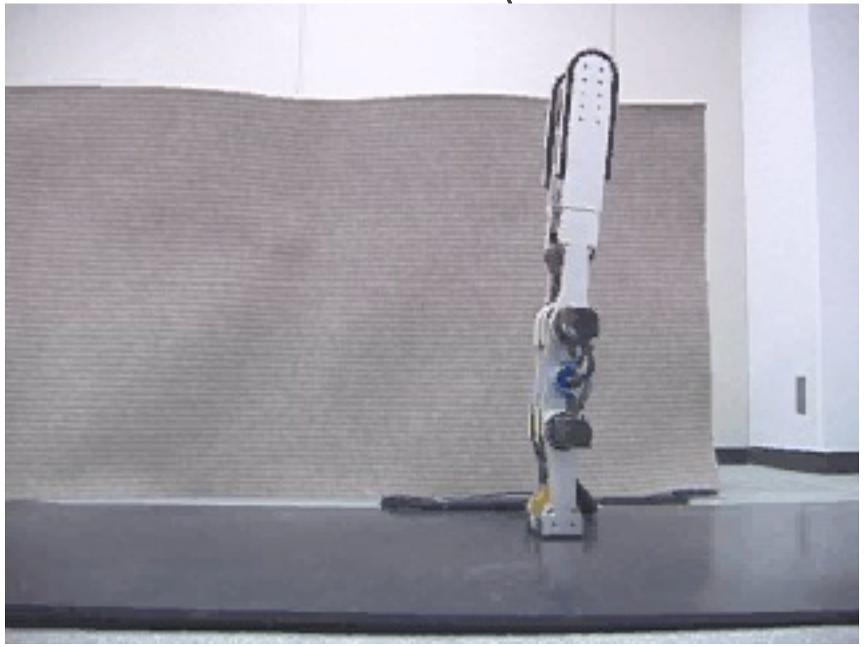
- N states, 2 actions



- if $r_2 >> r_1$, then better take a_2

Learning to Stand Up

(Morimoto & Doya, 2001)

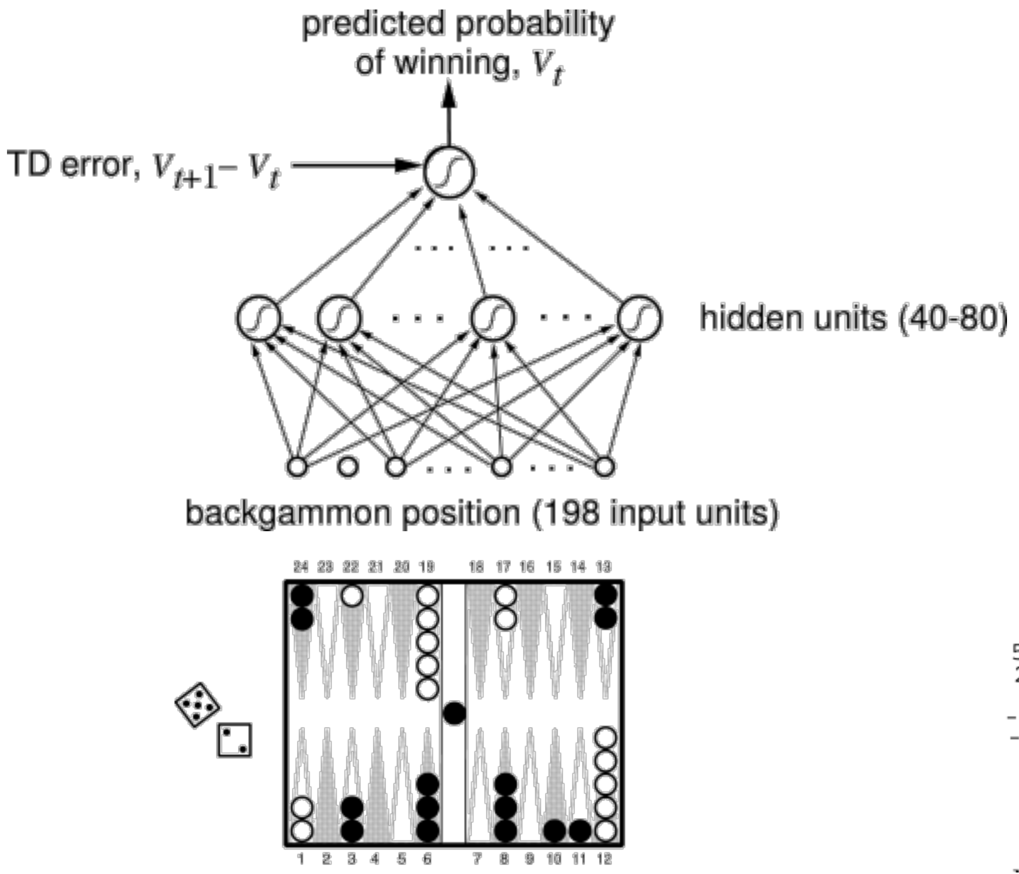


- Learning from reward and punishment
  - reward: height of the head
  - punishment: bump on the floor

TD Learning and Backprop

- TD Gammon

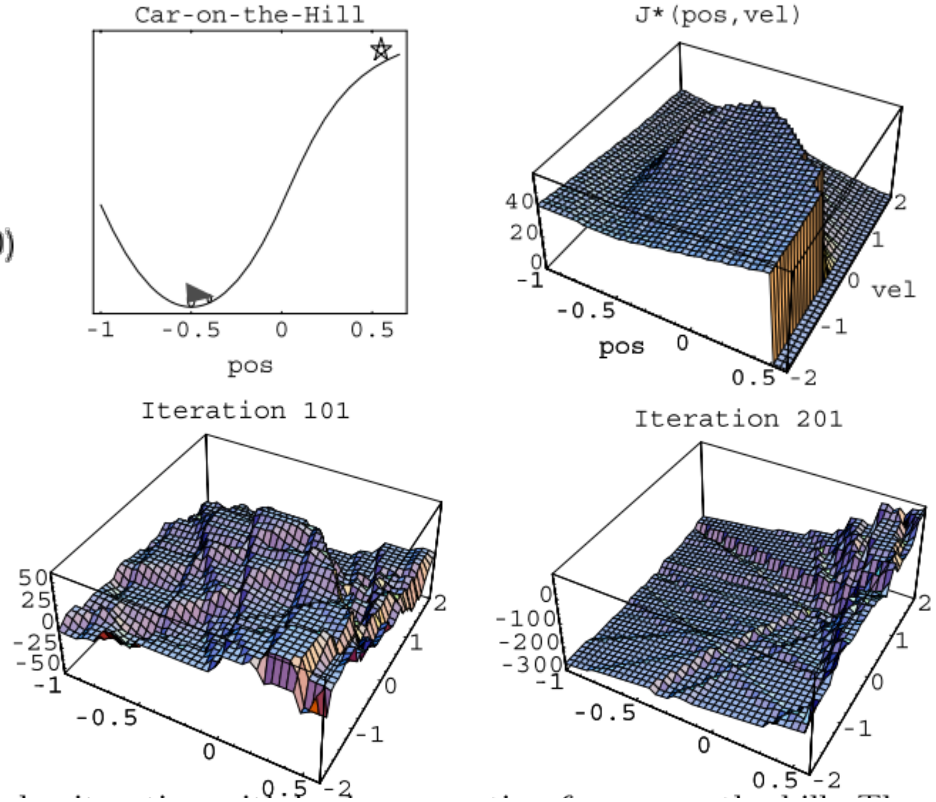
(Tesauro 1992, 1994)



- TD Learning can diverge

(Boyan & Moore, 1995)

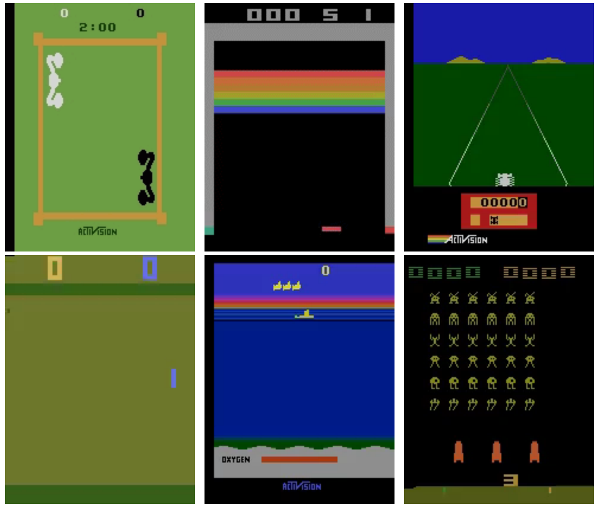
  - $\delta(t) = r(t) + \gamma\, V(s(t+1)) - V(s(t))$

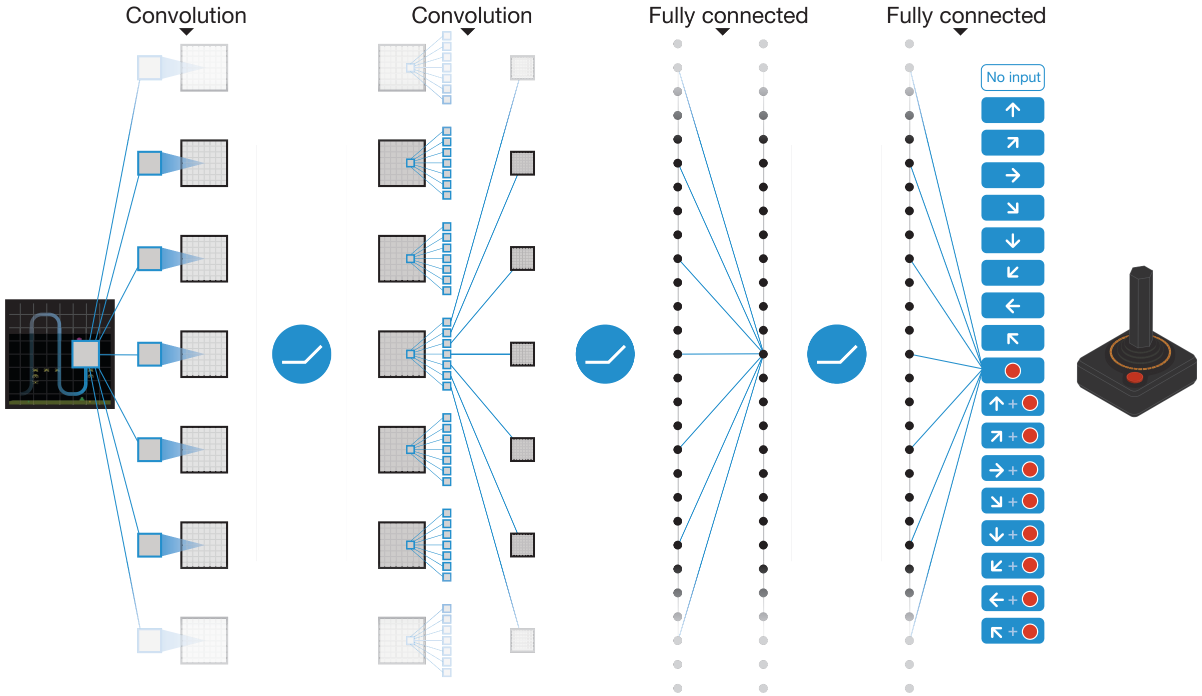


Deep Q-Network

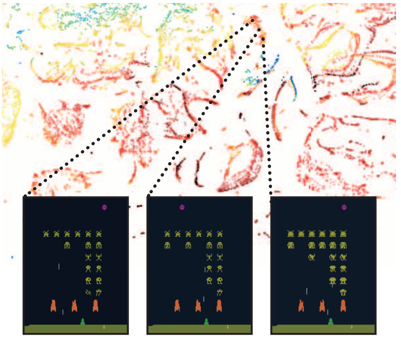
(Mnih et al. 2015)

- Game screen as input





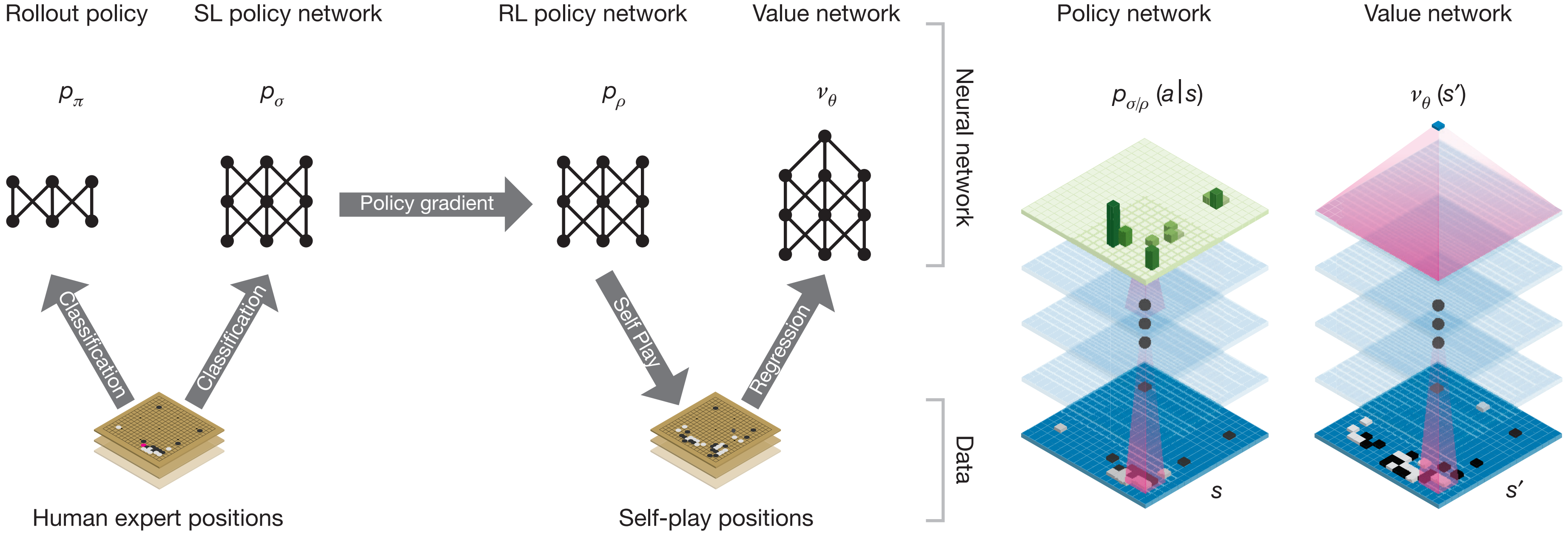
  - *Experience replay*
  - Fixing the *target network*
- DNN captures important features
  - human level in 29/49 Atari games



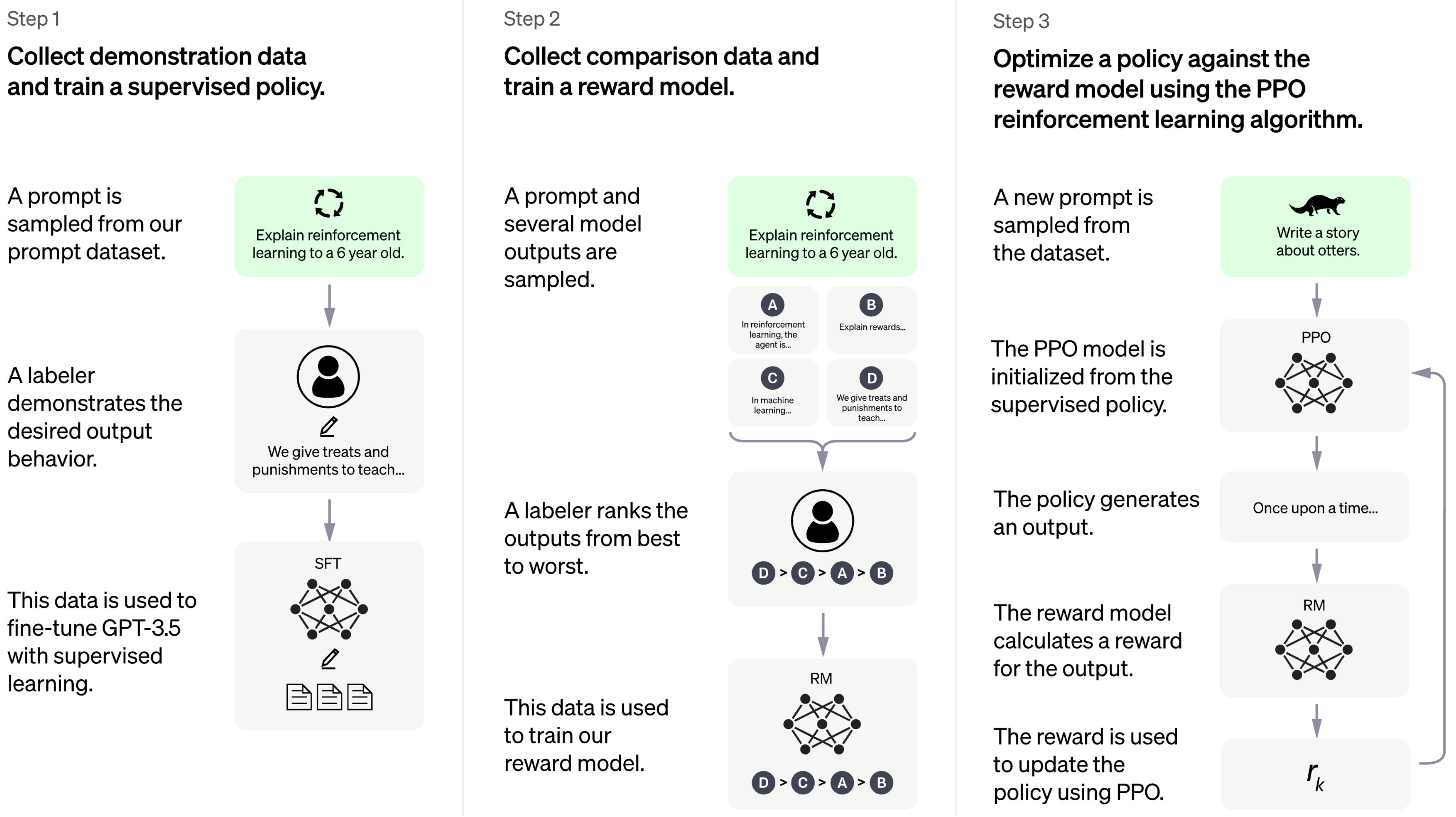
AlphaGo

(Silver et al., 2016)

- *Supervised learning* from play data
- *Reinforcement learning* by self-play
- *Representation learning* by deep neural networks
- Not too deep, wide tree search



ChatGPT



https://openai.com/blog/chatgpt

What is Bayesian Inference?

Joint probability: $P(X,Y) = P(X|Y)P(Y) = P(Y|X)P(X)$

Bayes theorem: $P(X|Y) = P(Y|X)P(X)/P(Y)$

Integrating prior belief and observation

X: unknown variable

Y: observation

- $P(X)$: prior probability of X
- $P(Y|X)$: probability of observing Y if X is true

  likelihood of X after observing Y
- $P(X|Y)$: posterior probability of X after observing Y

Posterior $\propto$ Prior belief x Likelihood by observation

- $P(Y) = \Sigma_X P(Y|X) P(X)$: marginal likelihood

Sunshine and Temperature

- X: weather Y: temperature

P(Y|X)	<20 degree	20 to 30 degree	>30 degree
Sunny	0.1	0.2	0.7
Cloudy	0.2	0.5	0.3
Rainy	0.5	0.4	0.1

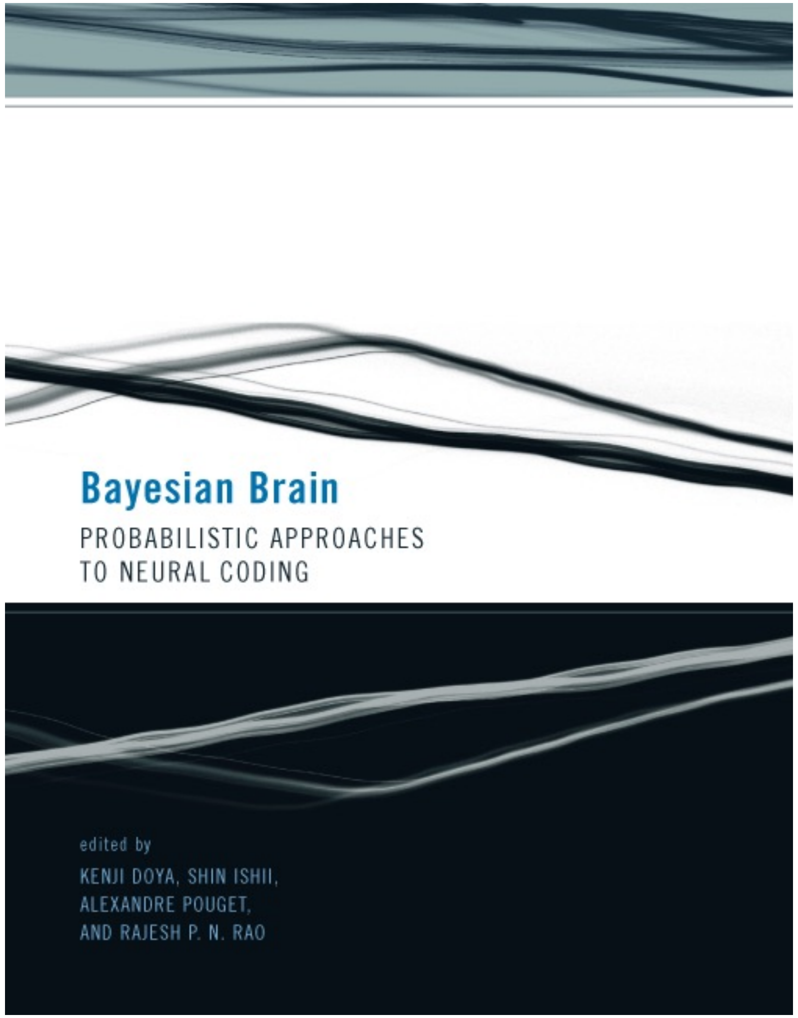
P(X)
0.5
0.3
0.2

- Temperature is 25 degree. What is the weather?
- Bayes theorem: $P(X|Y) = P(Y|X)P(X)/\Sigma_X P(Y|X)P(X)$
  - $P(s|Y) = P(Y|s)P(s)/\{P(Y|s)P(s)+P(Y|c)P(c)+P(Y|r)P(r)\}$
    $= 0.1/(0.1+0.15+0.08) = 0.1/0.33 \approx 0.3$

Bayesian Brain

Topics from OCNC 2004

- Kenji Doya, Shin Ishii
- Adrianne Fairhall
- Jonathan Pillow
- Barry Richmond
- Karl Friston
- Alex Pouget, Richard Zemel
- Peter Latham
- Tai Sing Lee
- David Knill
- Michael Shadlen
- Rajesh Rao
- Emanuel Todorov
- Konrad KÖrding



MIT Press, 2006

Dynamic Bayesian Inference

- Bayes rule: $P(x|y) = P(y|x)\,P(x)\,/\,P(y)$
  - sequential observation: $y_{1:t}=(y_1,\ldots,y_t)$
  - estimate hidden variable: $x_{1:t}=(x_1,\ldots,x_t)$
  - initial guess $P(x_1)$
- Dynamics model $P(x'|x)$
  - predictive prior

$$P(x_{t+1}|y_{1:t}) = \int P(x_{t+1}|x_t)P(x_t|y_{1:t})dx_t$$

- Observation model $P(y|x)$
  - new posterior

$$P(x_{t+1}|y_{1:t+1}) = P(y_{t+1}|x_{t+1})P(x_{t+1}|y_{1:t})\,/\,P(y_{1:t+1})$$

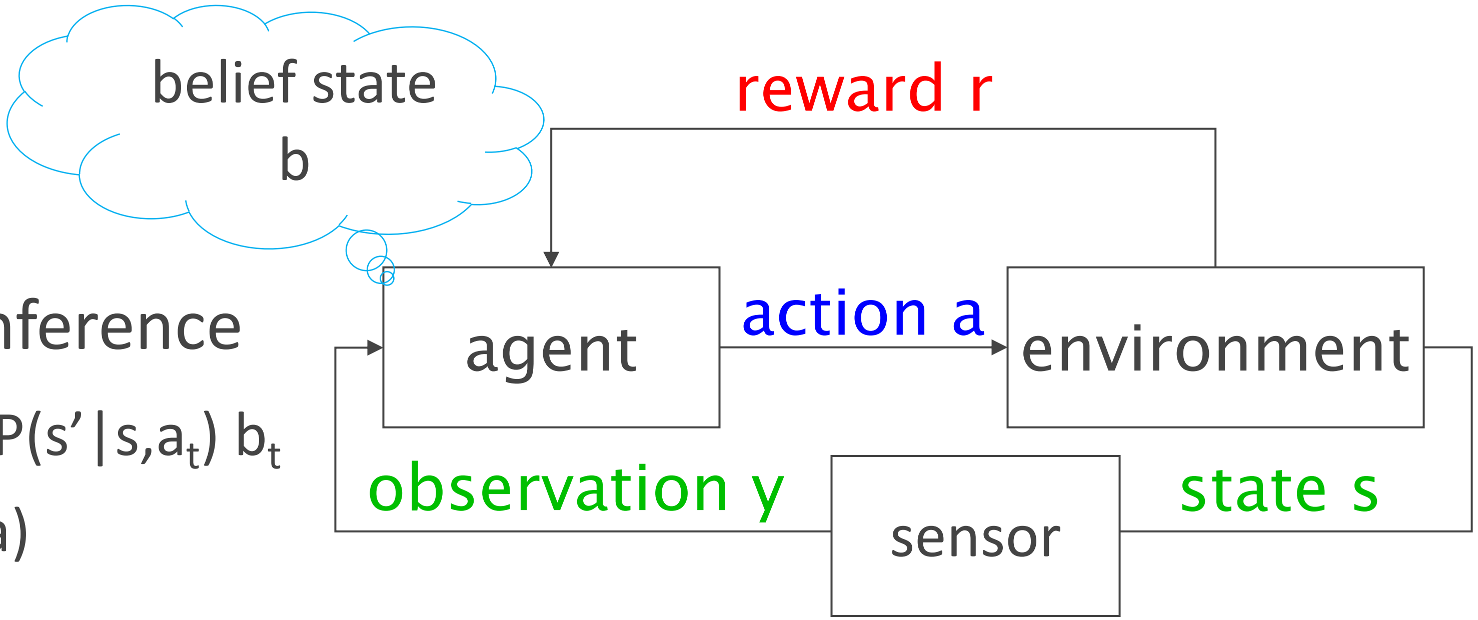
Partially Observable Markov Decision Process (POMDP)

- State is not fully observable
  - noise, delay, occlusion
- Update *belief state*:

  $b_t = P(s_t | y_{1:t}, a_{1:t-1})$
- Dynamic Bayesian inference

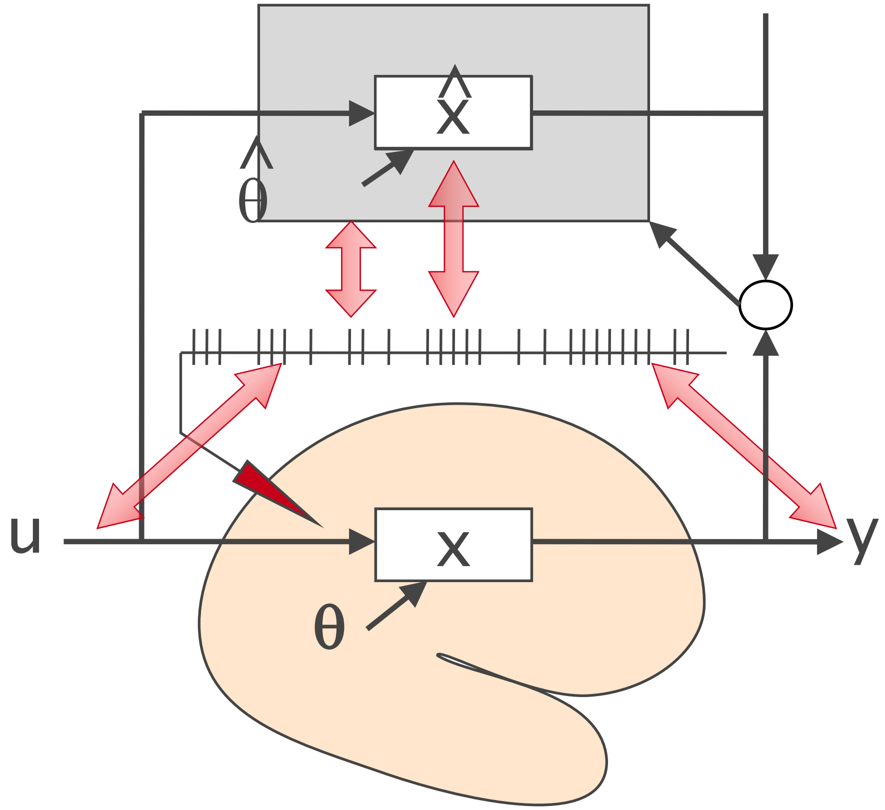
  $b_{t+1} \propto P(y_{t+1} | s') \sum_s P(s' | s, a_t)\, b_t$

  $Q(b,a) = \sum_s b_t\, Q(s,a)$



Model-based Neural Analysis

- Record and correlates with:
  - input u
  - output y
- internal state x
  - change by learning
- parameter θ
  - different in each session
- Run a dynamic model
  - estimate the internal variables
  - check correlation with recorded signal



ELSEVIER

Opinion

TRENDS in Neurosciences Vol.27 No.12 December 2004

Full text

The Bayesian brain: the role of uncertainty in neural coding and computation

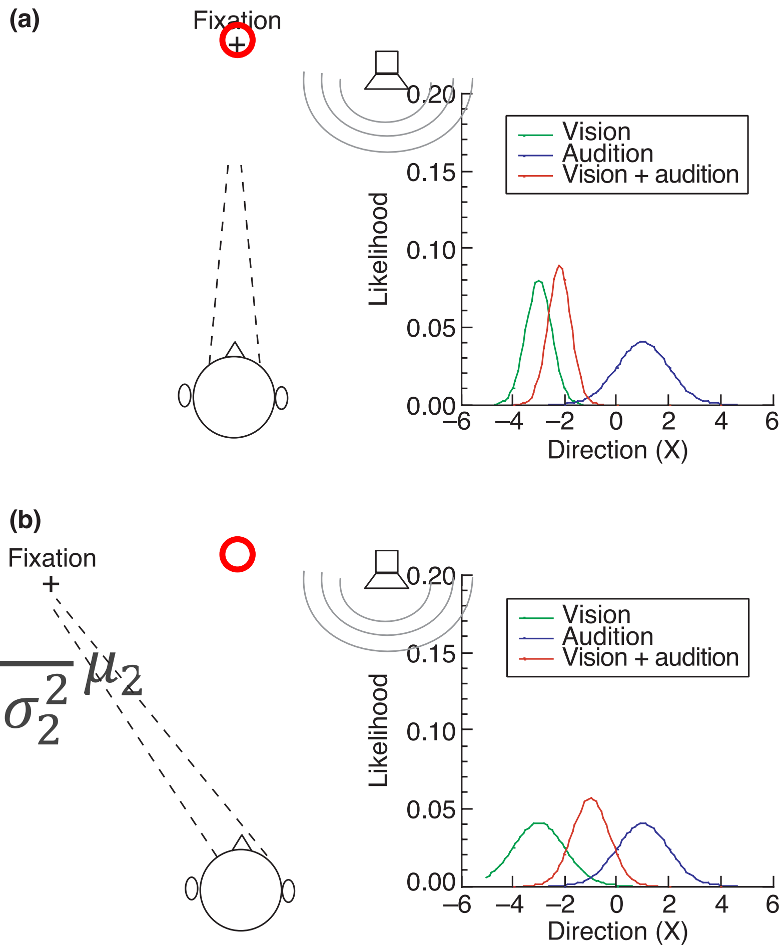
David C. Knill and Alexandre Pouget

- e.g. Sensory cue integration
  - p(X|V,A) ∝ p(V|X)p(A|X)p(X)
  - Gaussian noise, flat prior:

$$e^{-\frac{(x-\mu)^2}{2\sigma^2}} = e^{-\frac{(x-\mu_1)^2}{2{\sigma_1}^2}}\, e^{-\frac{(x-\mu_2)^2}{2{\sigma_2}^2}}$$

$$\mu = \frac{\sigma_2^2}{\sigma_1^2+\sigma_2^2}\mu_1 + \frac{\sigma_1^2}{\sigma_1^2+\sigma_2^2}\mu_2$$

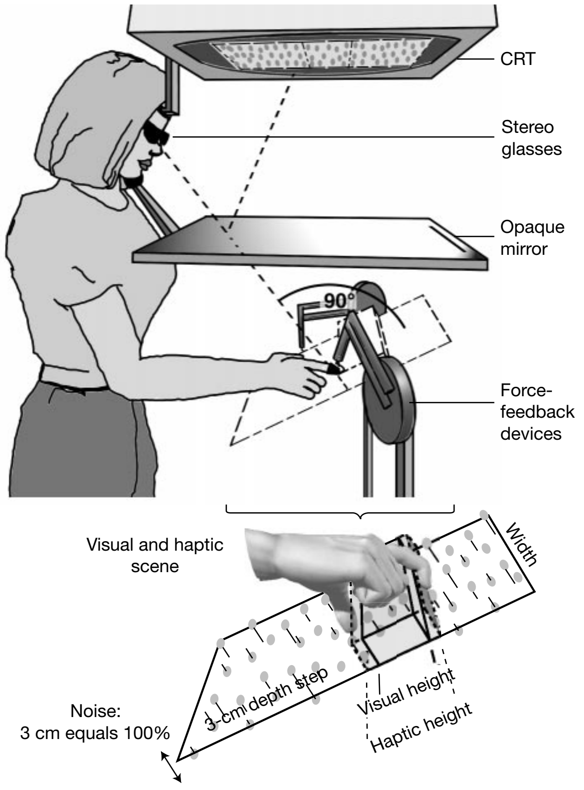
$$\sigma^2 = \frac{\sigma_1^2\sigma_2^2}{\sigma_1^2+\sigma_2^2}$$

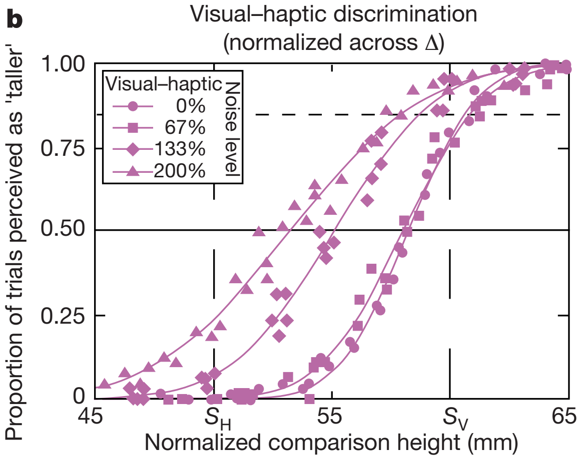


Multi-Sensory Integration

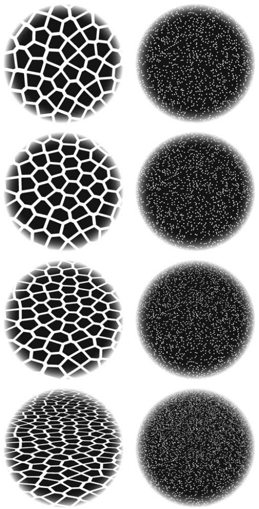
Humans integrate visual and haptic information in a statistically optimal fashion (2002, Nature)

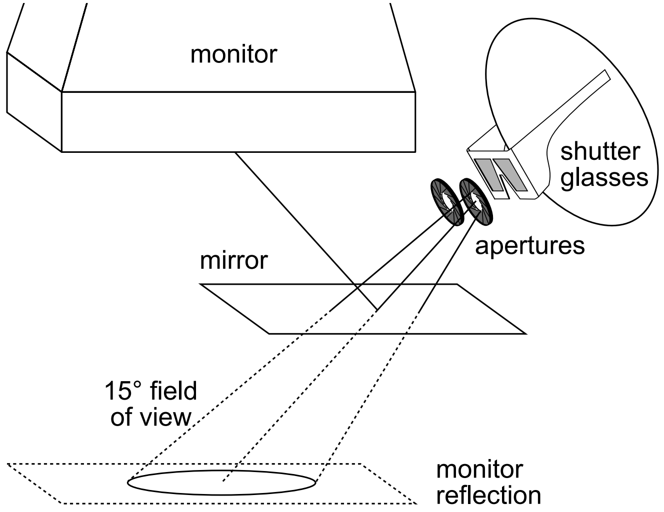
Marc O. Ernst* & Martin S. Banks

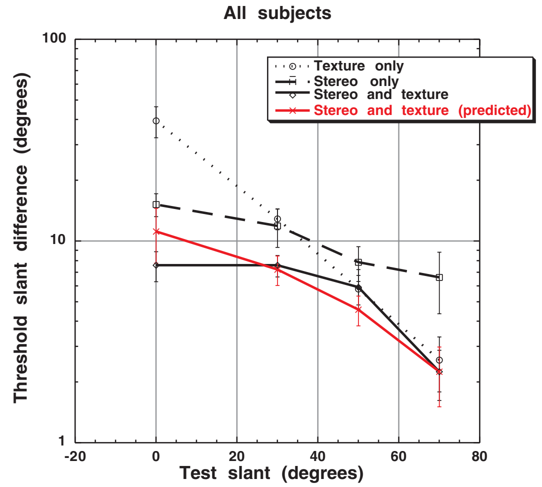




Knill & Saunders, (2003, Vision Research)



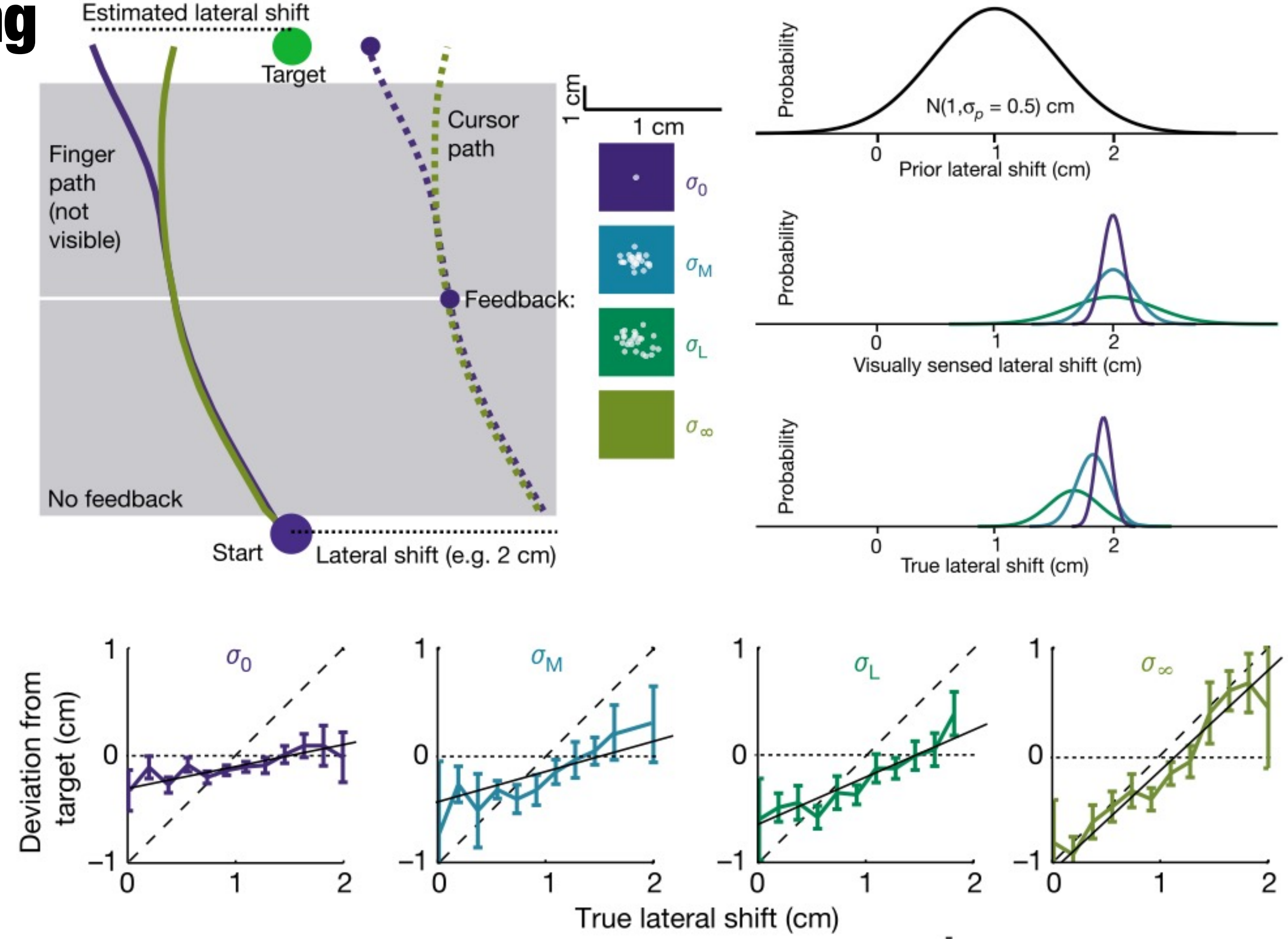




letters to nature (2004)

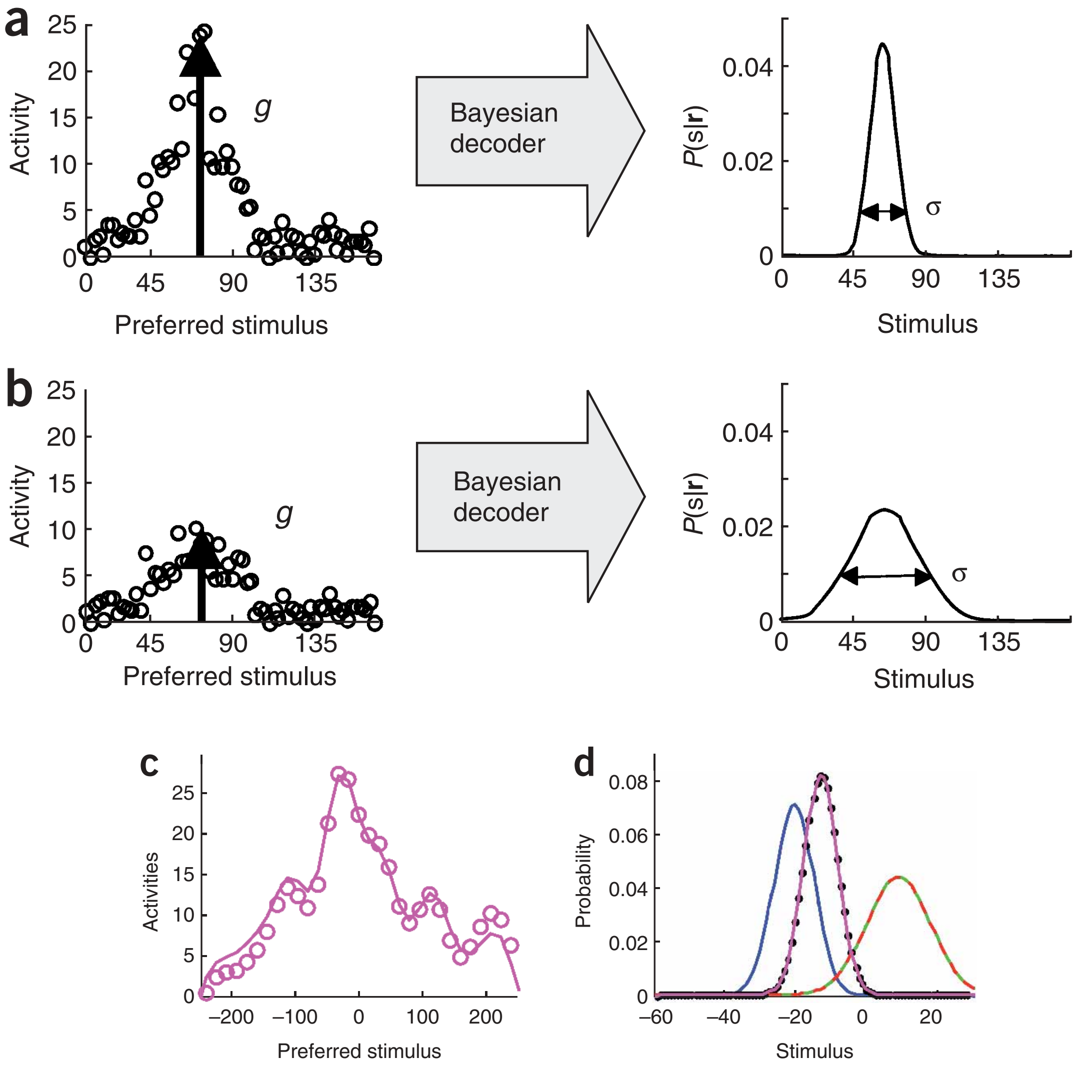
Bayesian integration in sensorimotor learning

Konrad P. Körding & Daniel M. Wolpert



Bayesian inference with probabilistic population codes

Wei Ji Ma[1,3], Jeffrey M Beck[1,3], Peter E Latham[2] & Alexandre Pouget[1] (2006, Nature Neuroscience)



Bayesian Model Selection

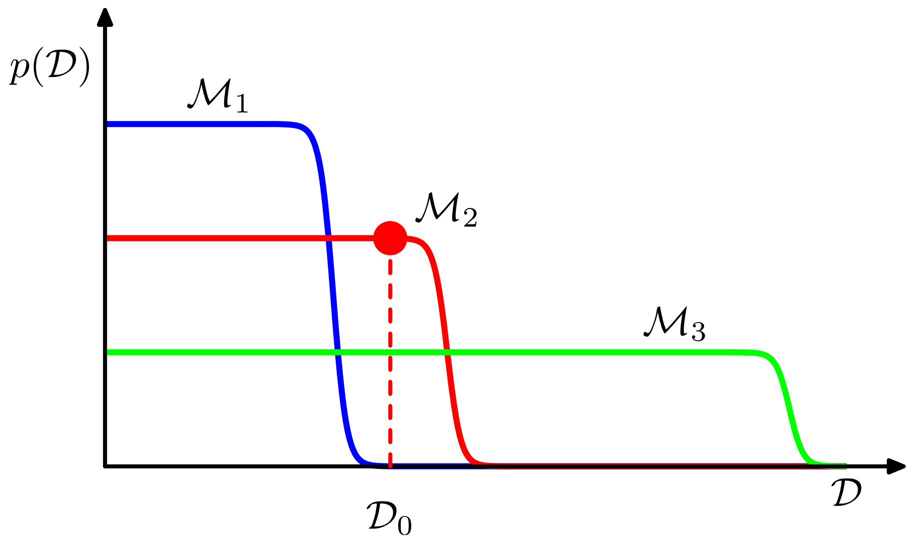
- Bayes rule: P(θ|Y) = P(Y|θ) P(θ) / P(Y)

- Denominator: marginal likelihood

  $P(Y)=\int P(Y|\theta)P(\theta)\, d\theta$

  - Measure of compatibility of model and data

- Too simple model
  - likelihood P(Y|θ) is low
- Too complex model
  - penalized by thin P(θ)
- ‘Evidence’ of model



Reinforcement Learning

- Predict reward: *value function*
  - $V(s) = E[\ r(t) + \gamma r(t+1) + \gamma^2 r(t+2)\ldots|\ s(t)=s]$
  - $Q(s,a) = E[\ r(t) + \gamma r(t+1) + \gamma^2 r(t+2)\ldots|\ s(t)=s,\ a(t)=a]$
- Select action
  - *greedy*: $a = \mathrm{argmax}\ Q(s,a)$
  - *Boltzmann*: $P(a|s) \propto \exp[\ \beta\ Q(s,a)]$
- Update prediction: *temporal difference* (*TD*) *error*
  - $\delta(t) = r(t) + \gamma V(s(t+1)) - V(s(t))$
  - $\Delta V(s(t)) = \alpha\ \delta(t)$
  - $\Delta Q(s(t),a(t)) = \alpha\ \delta(t)$

How to implement these steps?

How to tune these parameters?

Basal Ganglia

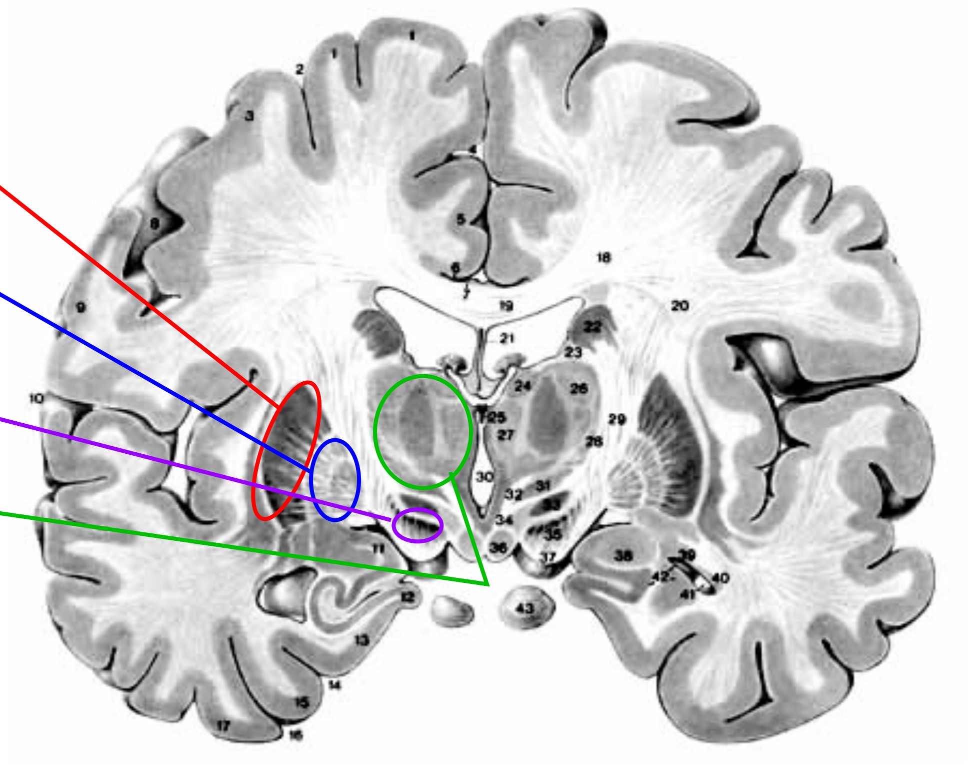
- Locus of Parkinson’s and Huntington’s diseases

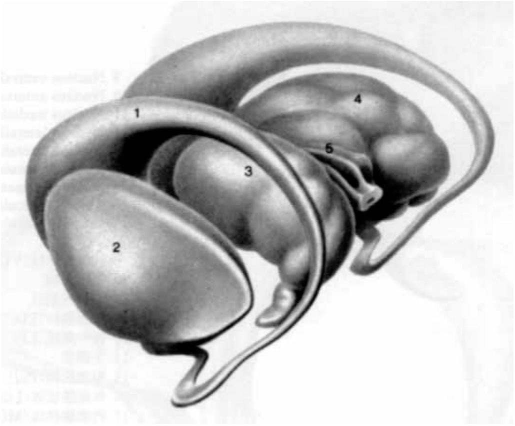
Striatum

Globus Pallidus

Substantia Nigra

Thalamus

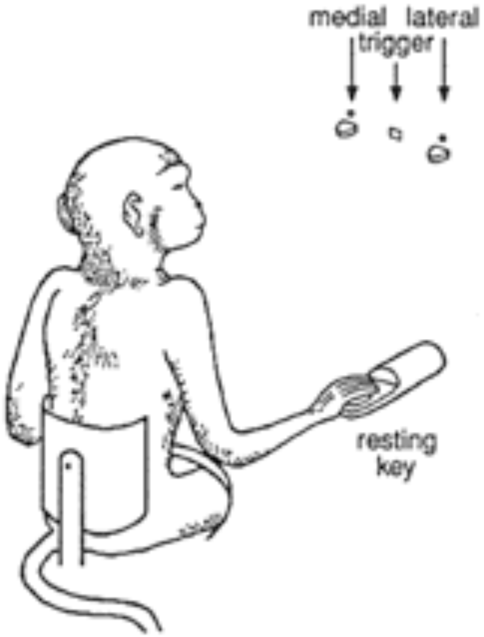


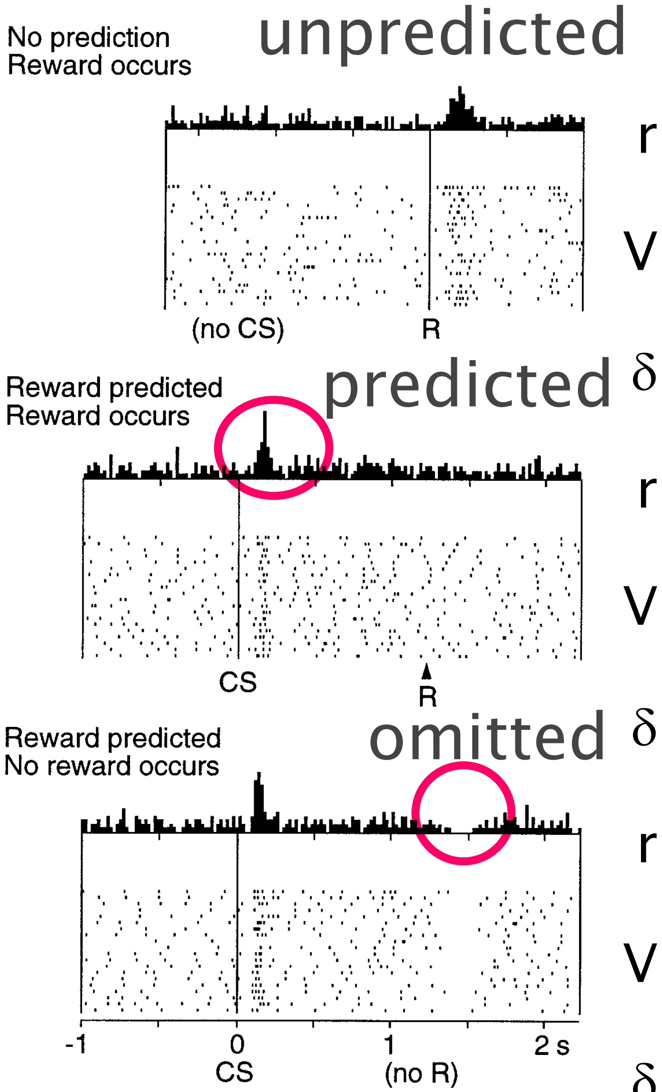


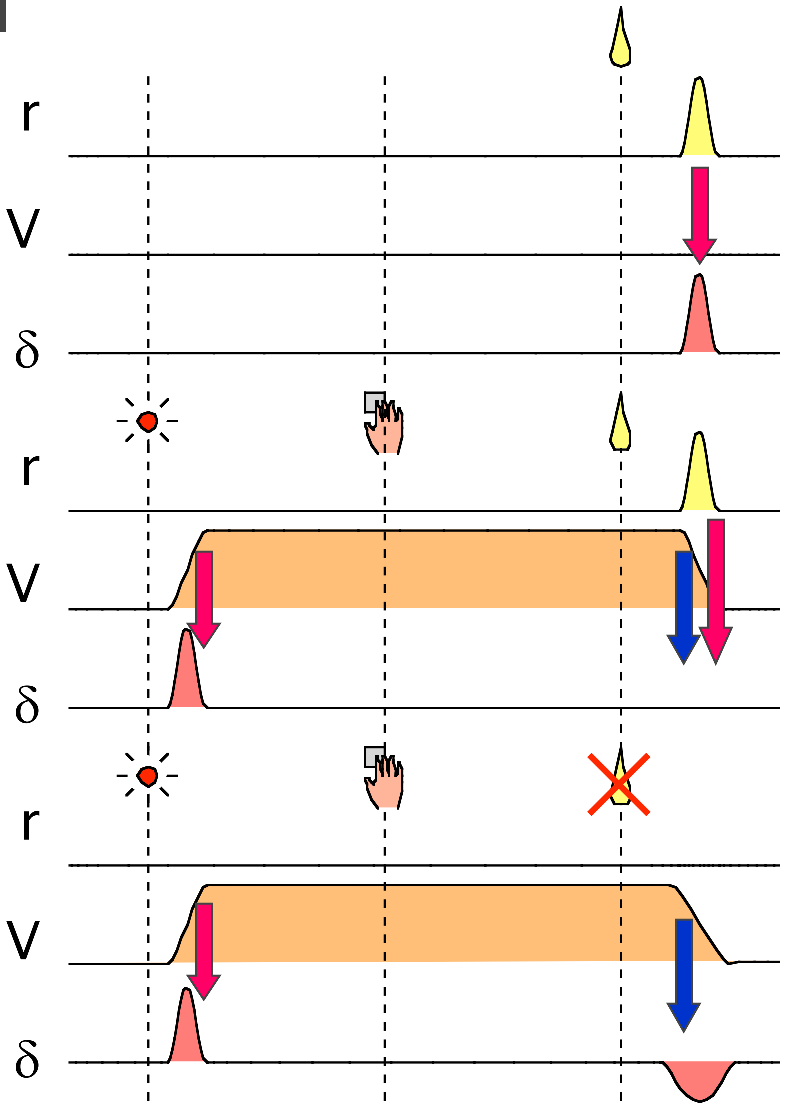
- What is their normal function??

Dopamine Neurons Code TD Error

$$\delta(t) = r(t) + \gamma V(s(t+1)) - V(s(t))$$



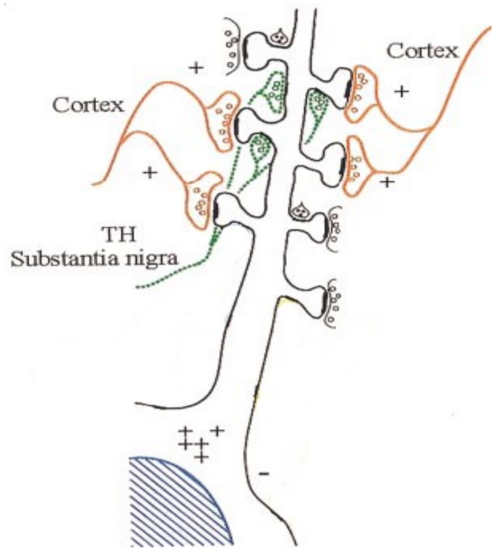


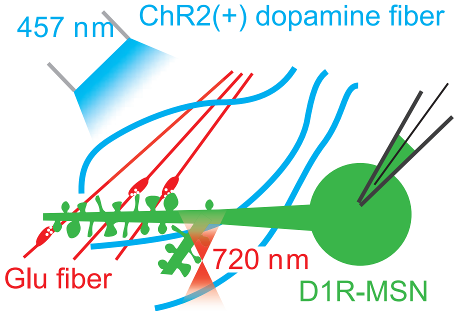


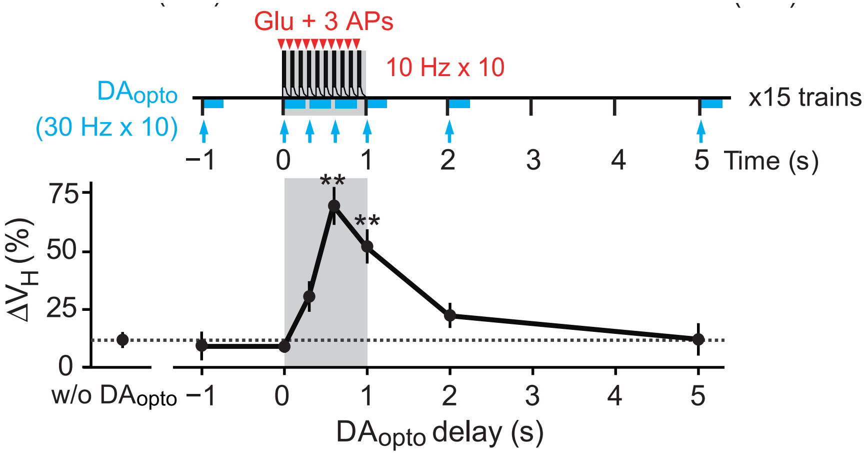
(Schultz et al. 1997)

Dopamine-dependent Plasticity

- Medium spiny neurons in striatum
  - glutamate from cortex
  - dopamine from midbrain
- Three-factor learning rule (Wickens et al.)
  - cortical input + spike → LTD
  - cortical input + spike + dopamine → LTP
  - input x output x reward
- Time window of plasticity
  (Yagishita et al., 2014)

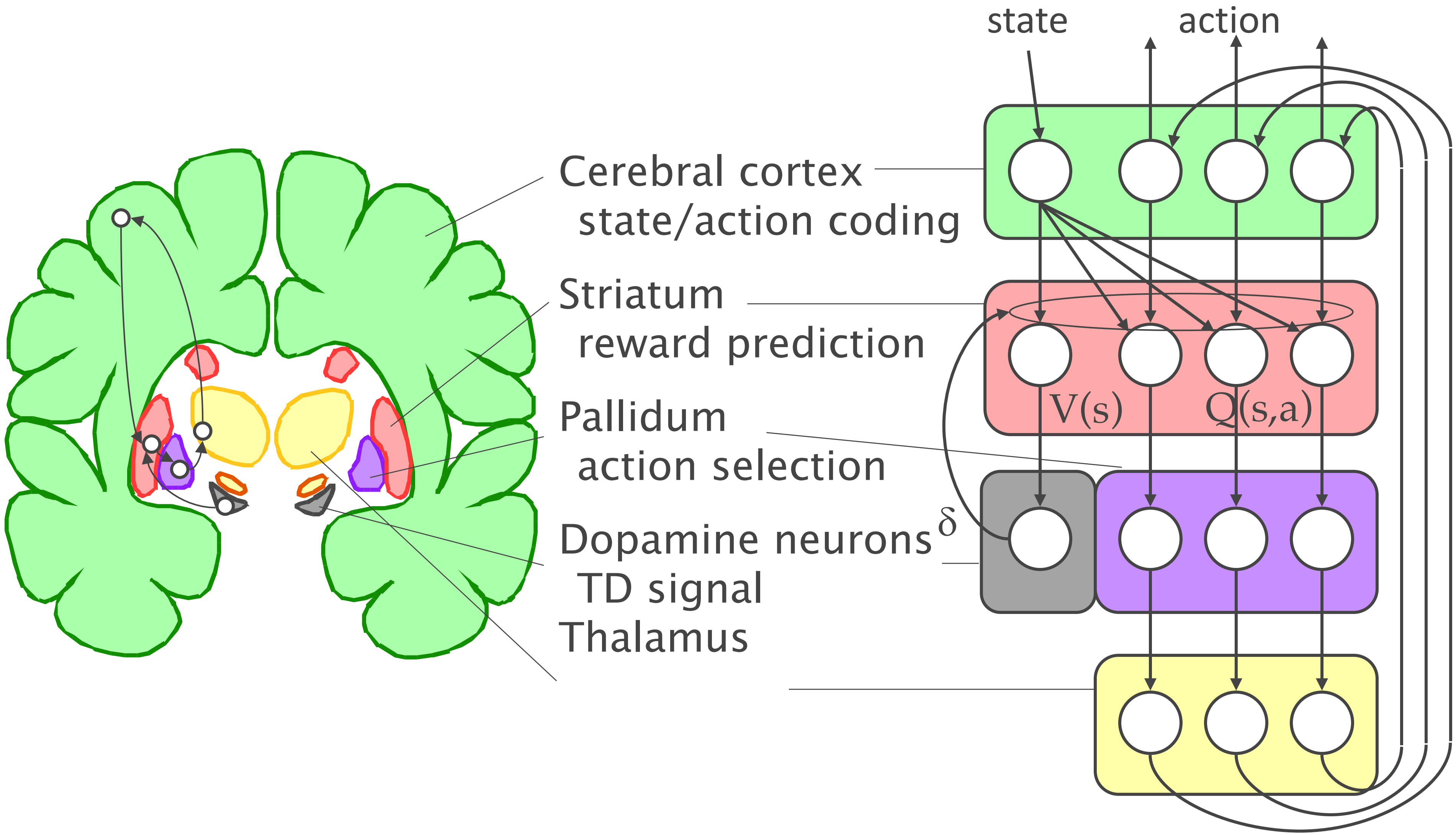






Basal Ganglia for Reinforcement Learning?

(Doya 2000, 2007)

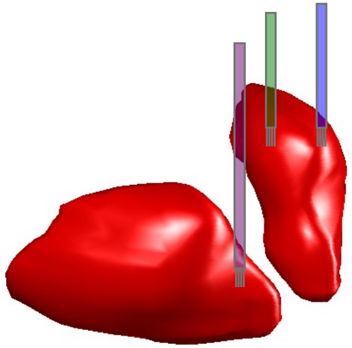


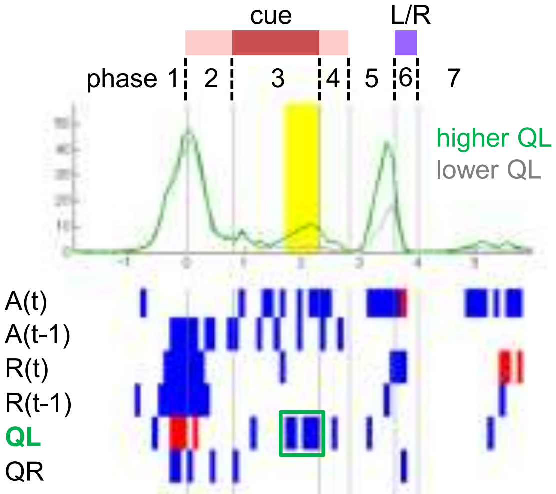
Distinct Neural Representation in the Dorsolateral, Dorsomedial, and Ventral Parts of the Striatum during Fixed- and Free-Choice Tasks

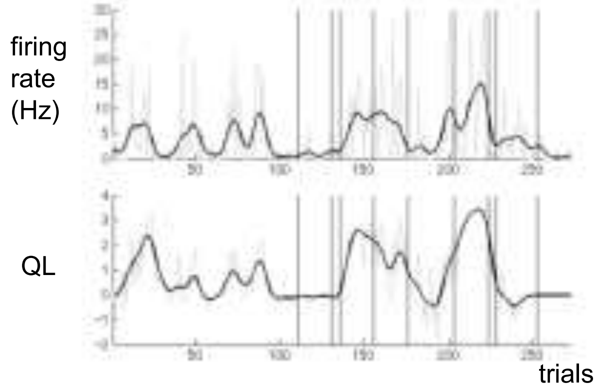
Makoto Ito and Kenji Doya

The Journal of Neuroscience, 2015

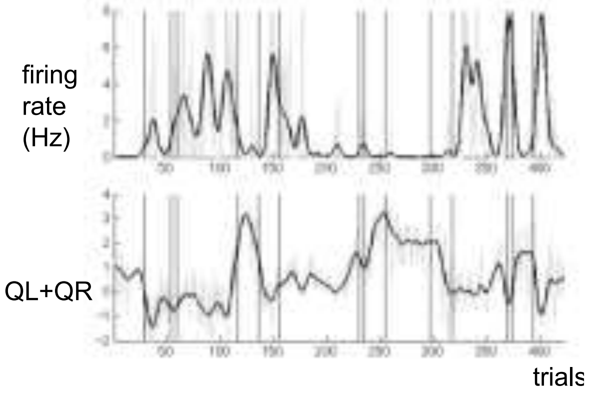
Left Center Right







- Dorsolateral
  - movements
- Dorsomedial
  - action value
- Ventral
  - state value



Generalized Q-learning Model

(Ito & Doya, 2009)

- Action selection

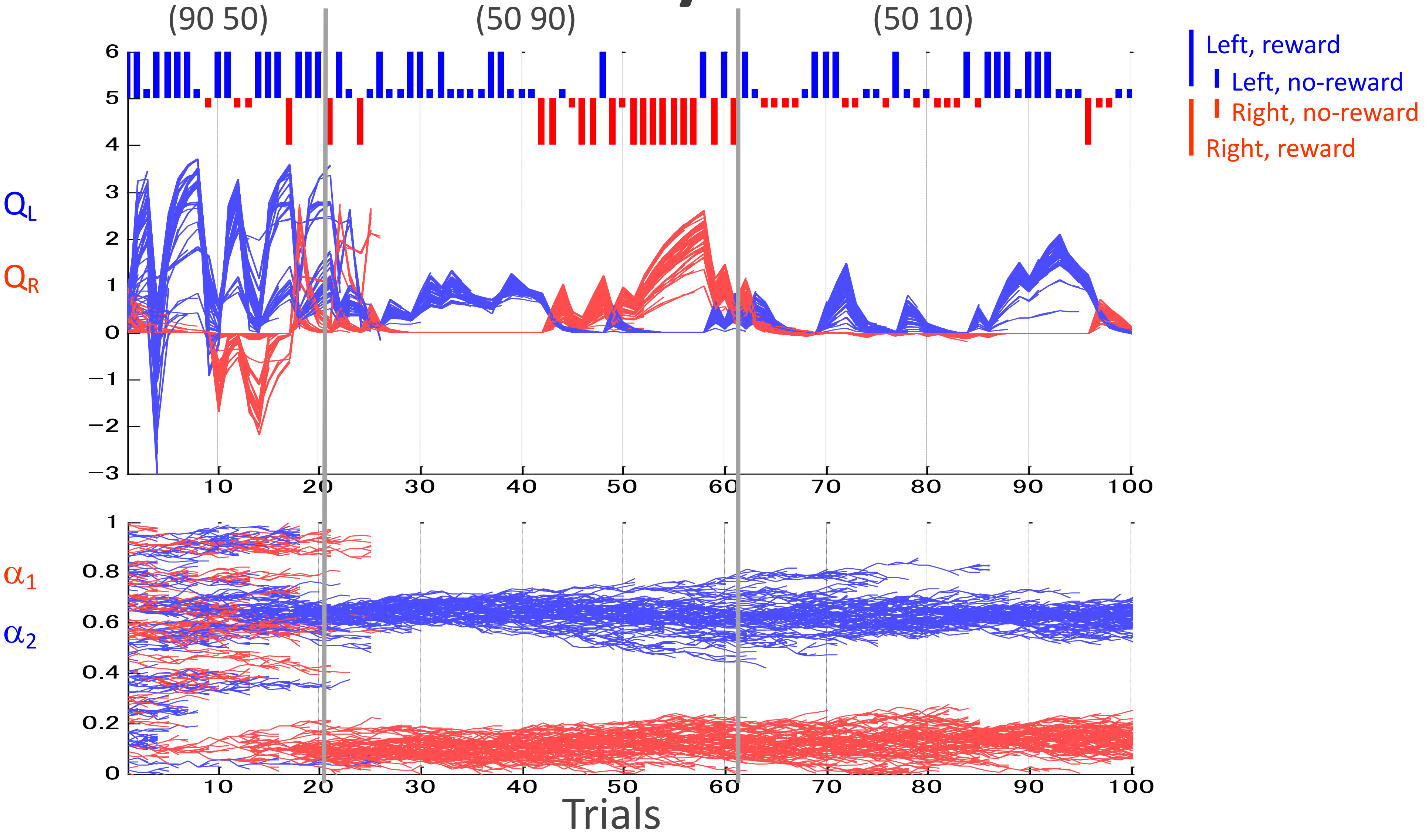
$$P(a(t)=L) = \exp Q_L(t)/(\exp Q_L(t)+\exp Q_R(t))$$

- Action value update: $i \in \{L,R\}$

$$Q_i(t+1) = \begin{cases} (1-\alpha_1)Q_i(t) + \alpha_1\kappa_1 & \text{if } a(t)=i,\ r(t)=1 \\ (1-\alpha_1)Q_i(t) - \alpha_1\kappa_2 & \text{if } a(t)=i,\ r(t)=0 \\ (1-\alpha_2)Q_i(t) & \text{if } a(t)\neq i,\ r(t)=1 \\ (1-\alpha_2)Q_i(t) & \text{if } a(t)\neq i,\ r(t)=0 \end{cases}$$

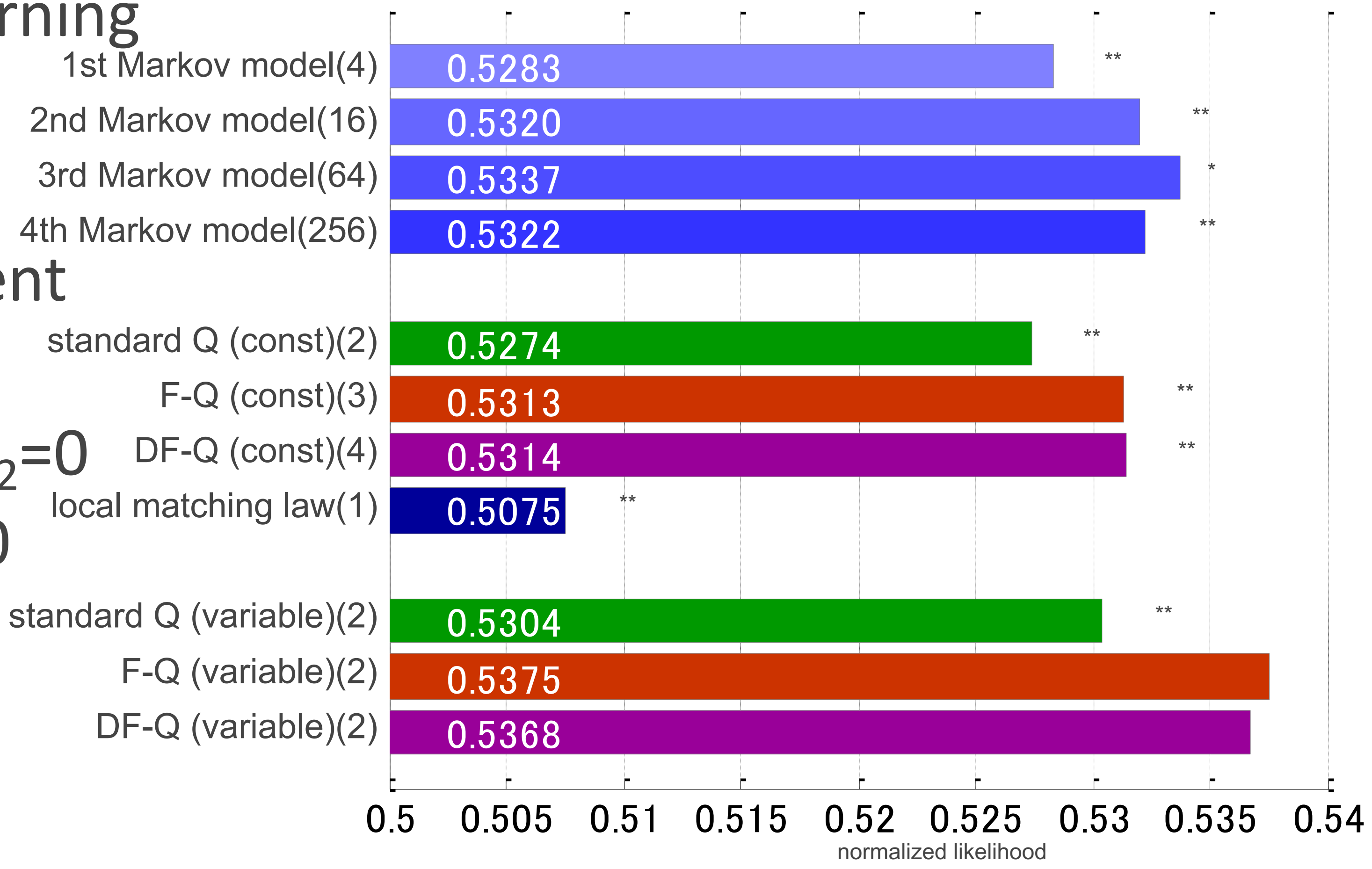
- Parameters
  - α_1: learning rate
  - α_2: forgetting rate
  - κ_1: reward reinforcement
  - κ_2: no-reward aversion

Estimation by Particle Filter



Model Fitting

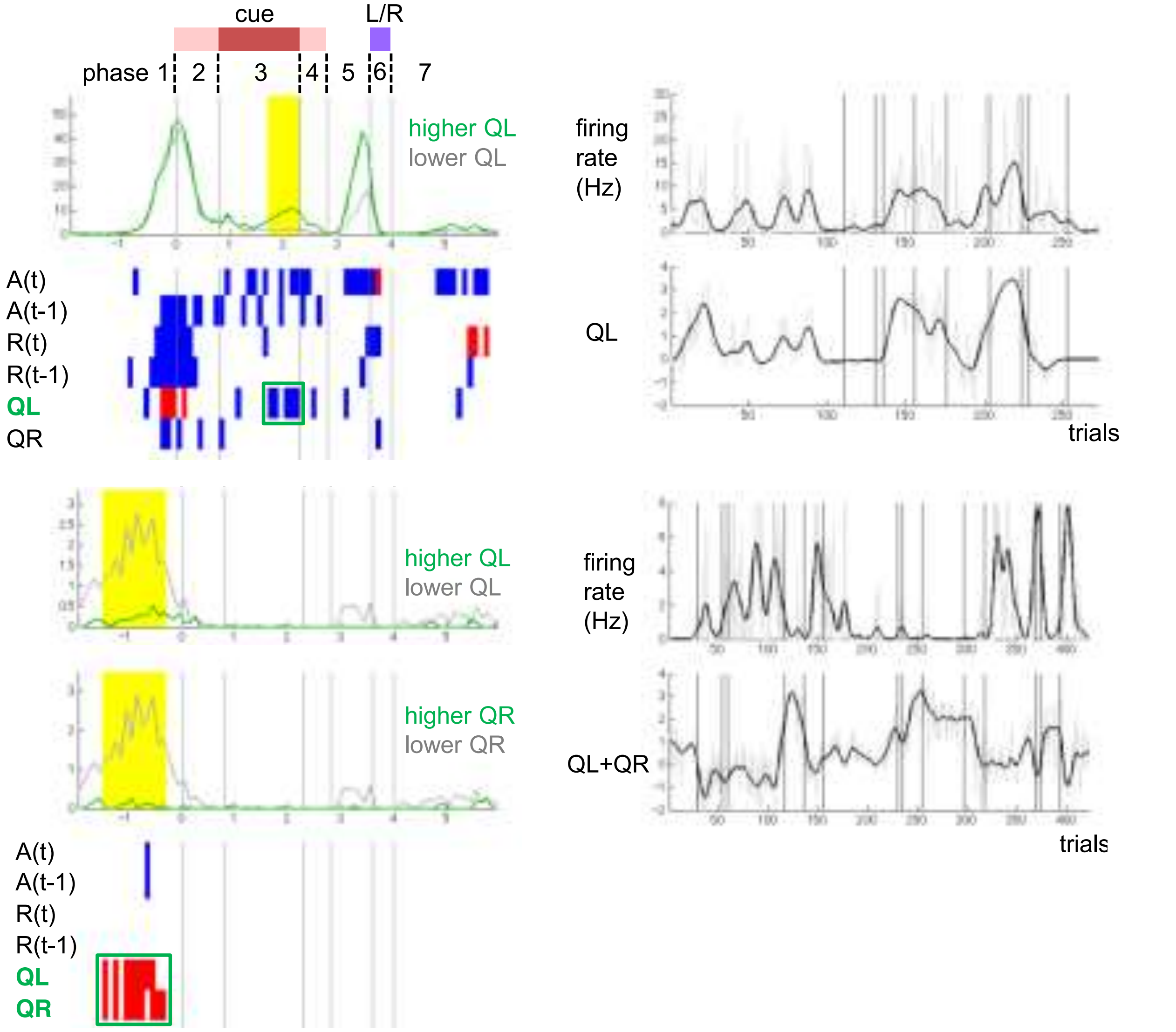
- Generalized Q learning
  - α_1: learning
  - α_2: forgetting
  - κ_1: reinforcement
  - κ_2: aversion
  - standard: $\alpha_2=\kappa_2=0$
  - forgetting: $\kappa_2=0$



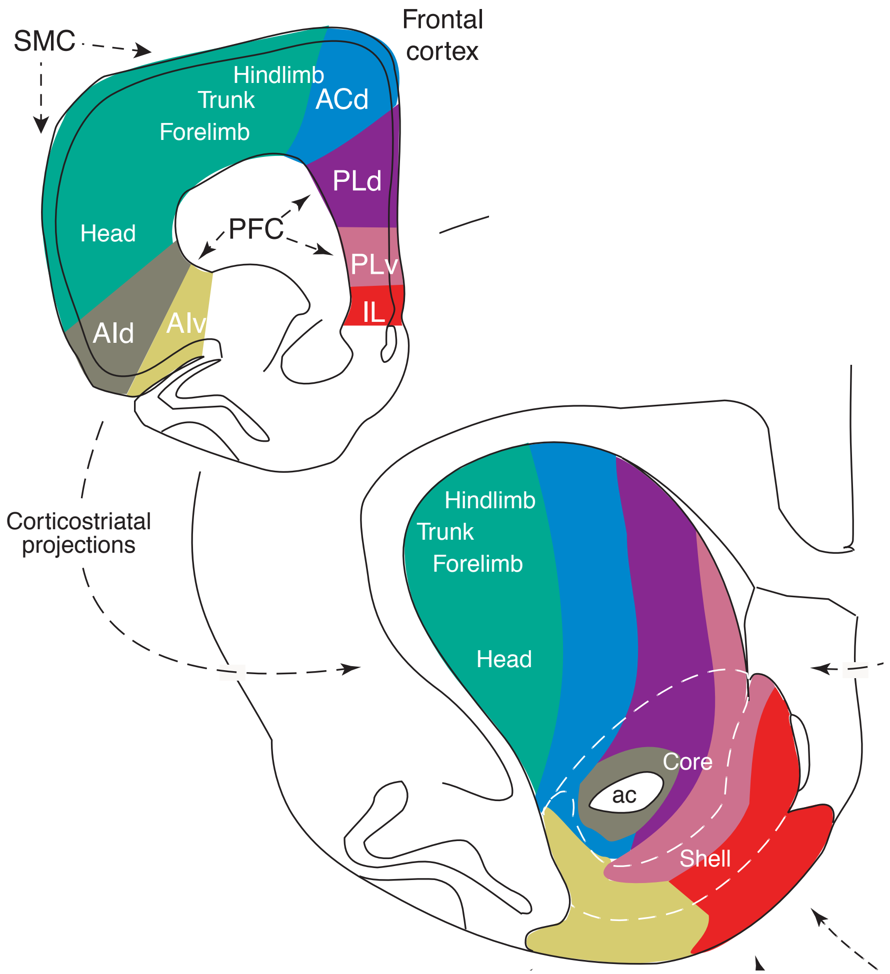
Action/State Value Coding Neurons

(Ito & Doya, 2015, JNS)

- Action value
  - DLS
  - DMS
- State value
  - VS



Hierarchy in Cortico-Striatal Network



(Voorn et al., 2004)

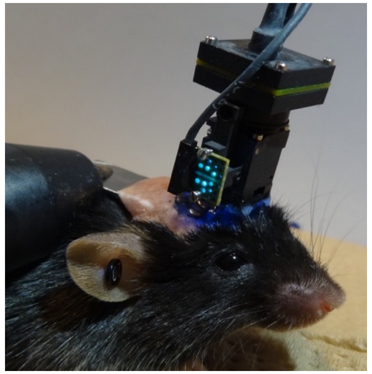
- Dorsolateral striatum: motor
  - early action coding
  - what motor action?
- Dorsomedial striatum: cognitive
  - choice action value
  - which goal?
- Ventral striatum: motivational?
  - state value
  - whether worth doing?

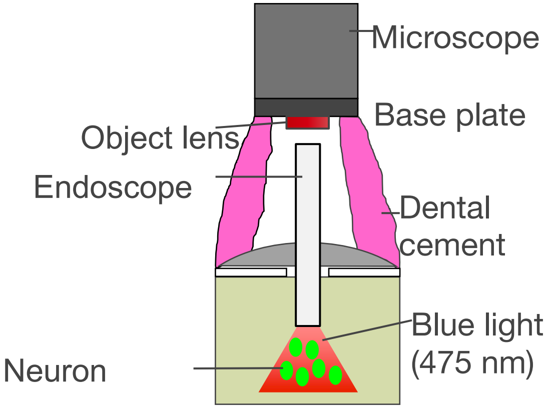
eNeuro (2018)

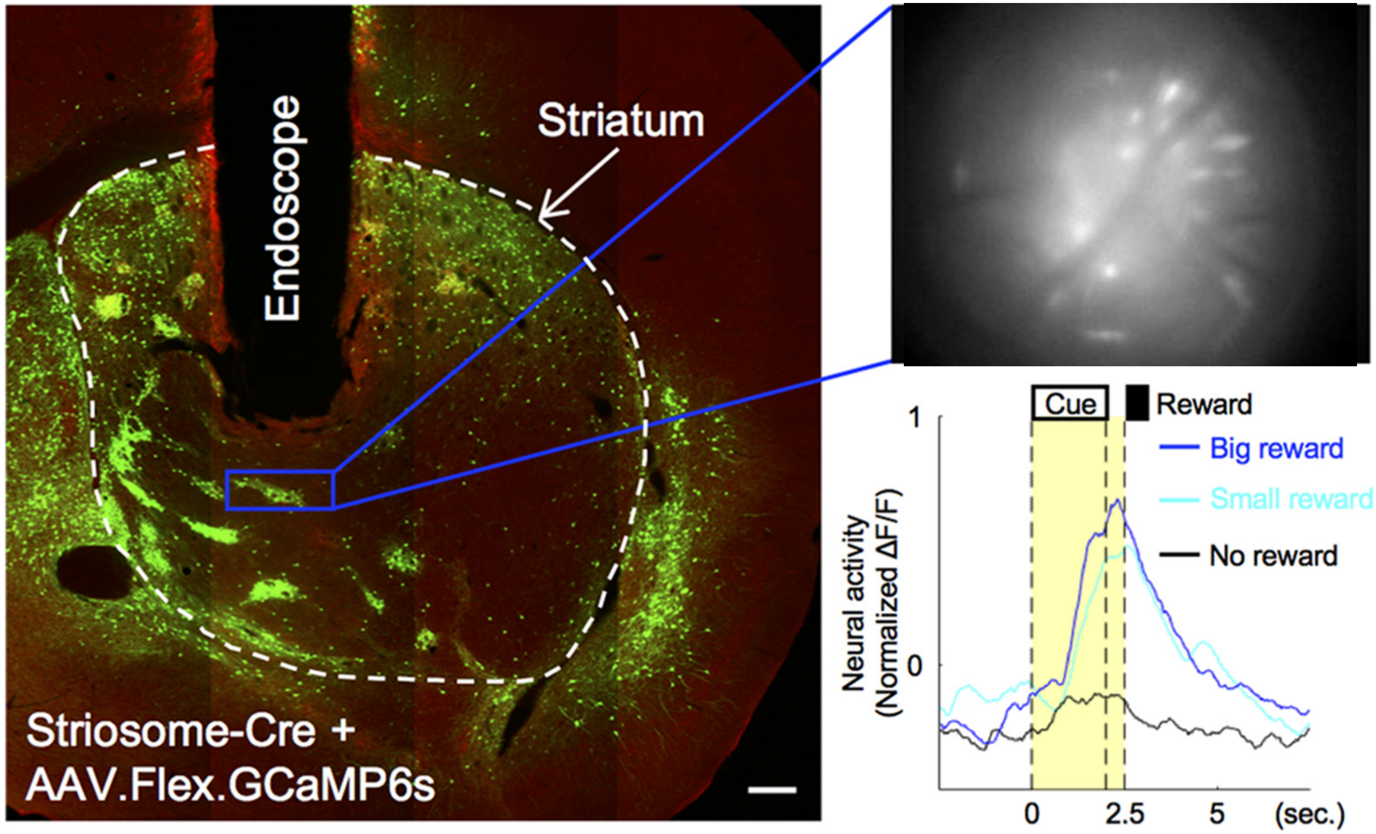
Reward-Predictive Neural Activities in Striatal Striosome Compartments

Tomohiko Yoshizawa,[1] Makoto Ito,[1,2] and Kenji Doya[1]

Imaging striosome neuron activity by endoscope



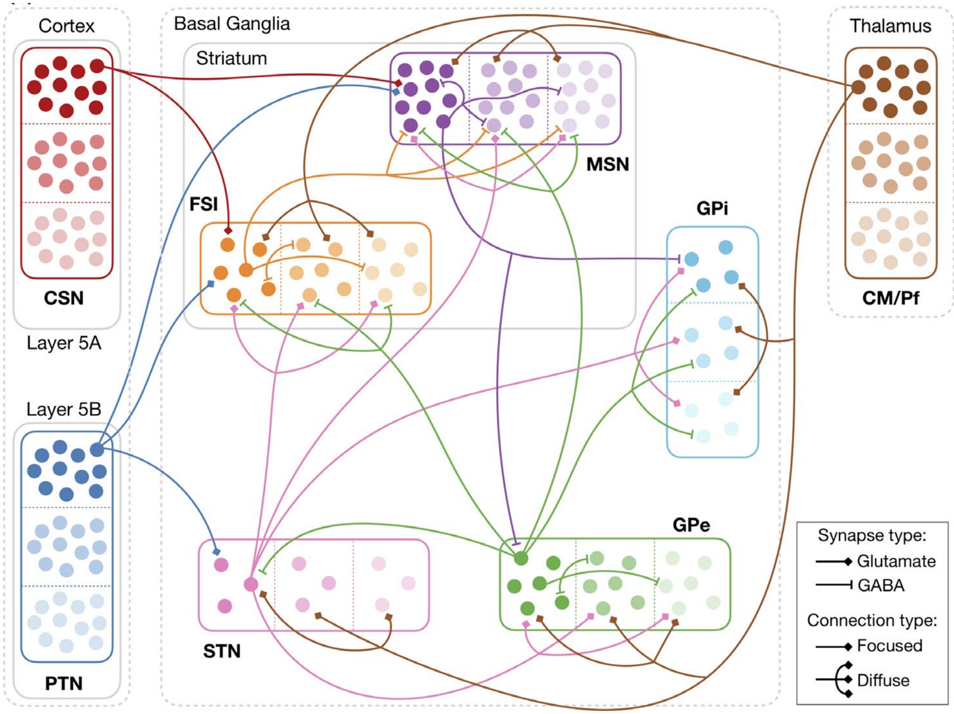




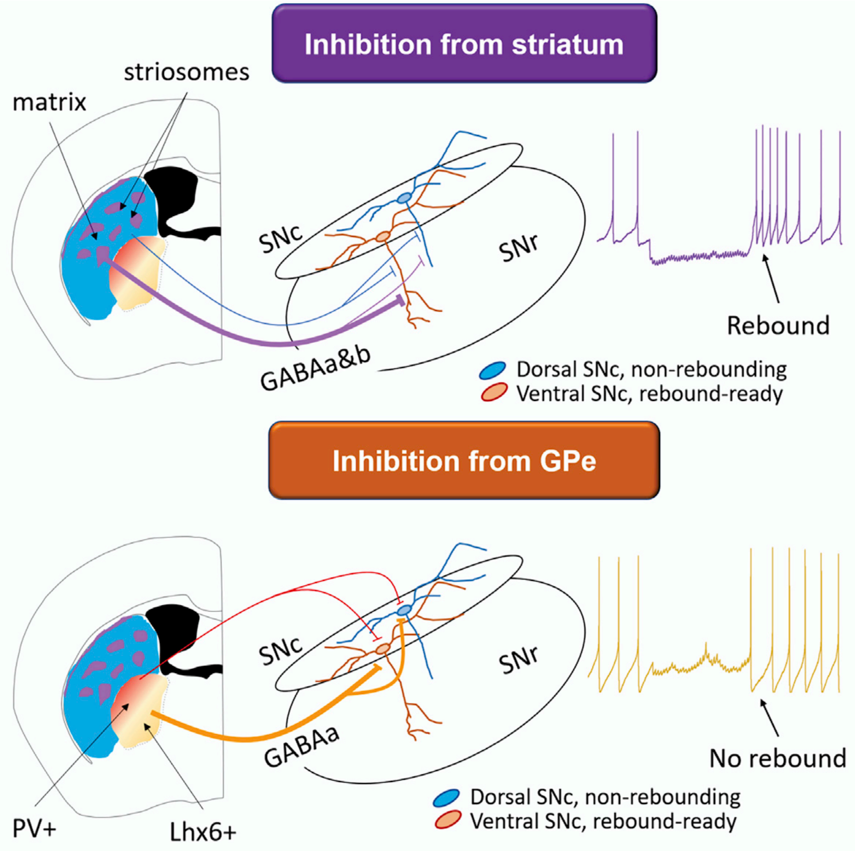
Open Questions

Parallel, multi-inhibitory pathways

(Girard et al. 2020)



TD like response of dopamine neurons



(Evans et al. 2020)

Amygdala, Hippocampus, Cerebellum,...

Model-free/Model-based Strategies

Model-free

- No knowledge of the world
- Learn values by experience
  - state–action–reward
- Act and then learn

Simple, but slow learning

Model-based

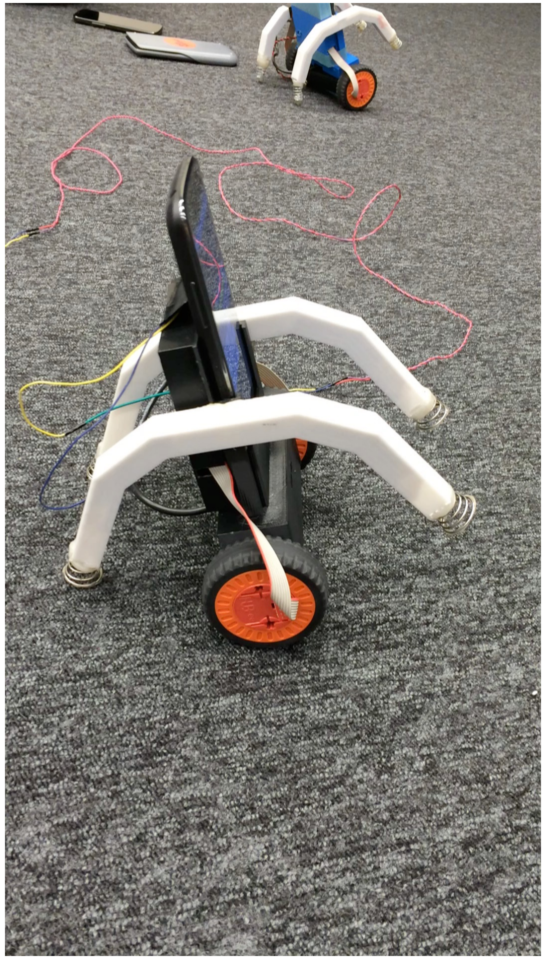
- Learn prediction model:
  - state, action $\rightarrow$ new state
- Internal simulation
  - estimate current state
  - plan future actions
- Predict and then act

Flexible, but heavy load

Bounce Up and Balance by PILCO

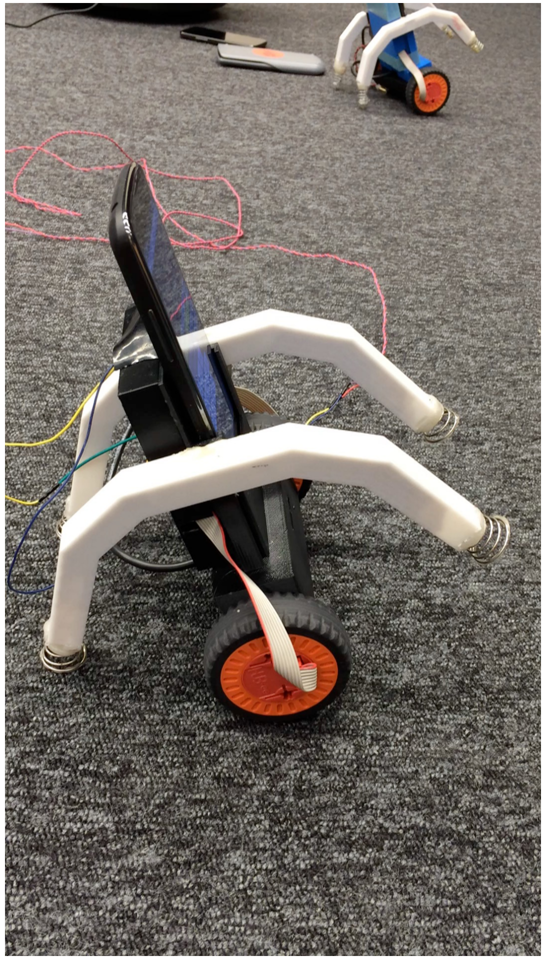
(Paavo Parmas)

1st try



2nd try

8th try



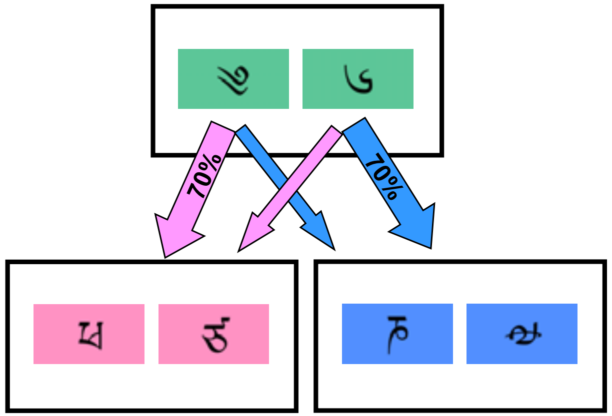
Mental Simulation

Brain's process using
an action-dependent state transition model
$s'=f(s,a)$ or $P(s'|s,a)$

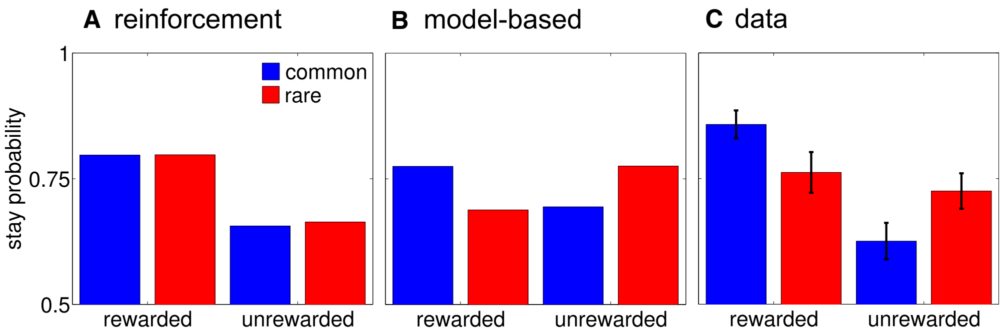
- Estimate the present from past state/action
  - perception under noise/delay/occlusion
- Predicting the future
  - model-based decision, action planning
- Imagining in a virtual world
  - thinking, language, science,…

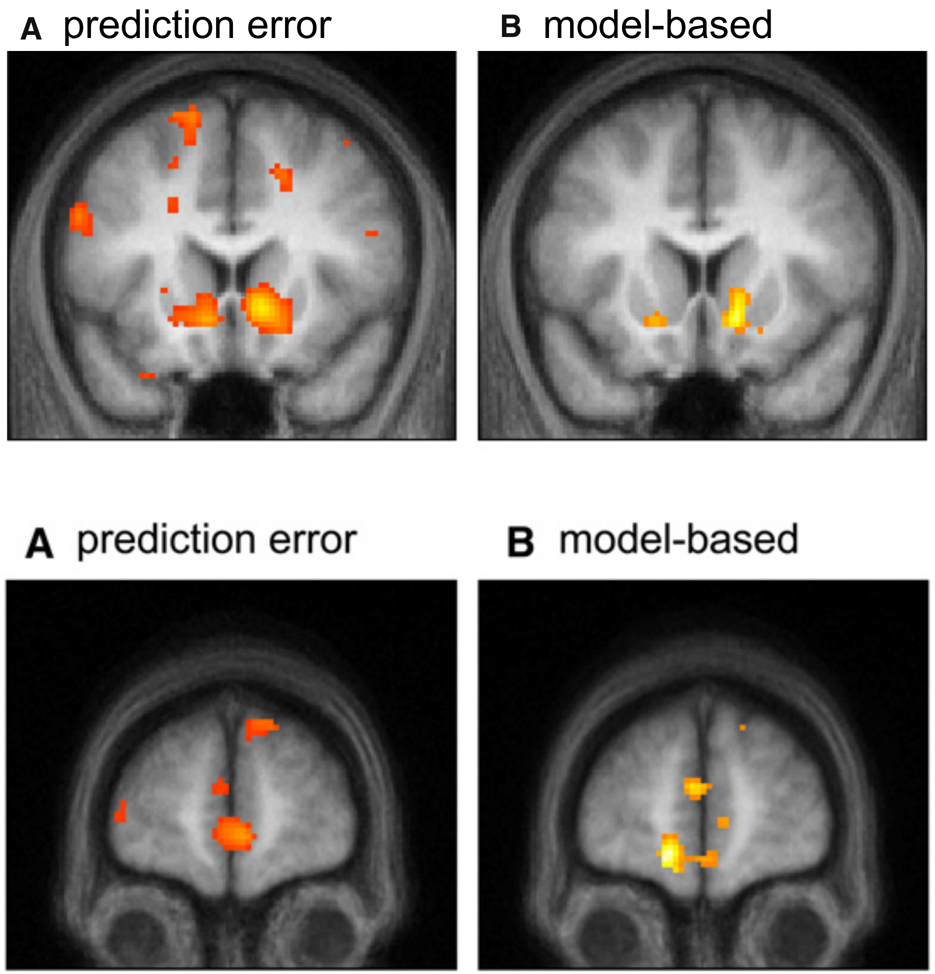
Model-free and Model-based Choice

(Daw et al. 2011)



- choice after rare transition

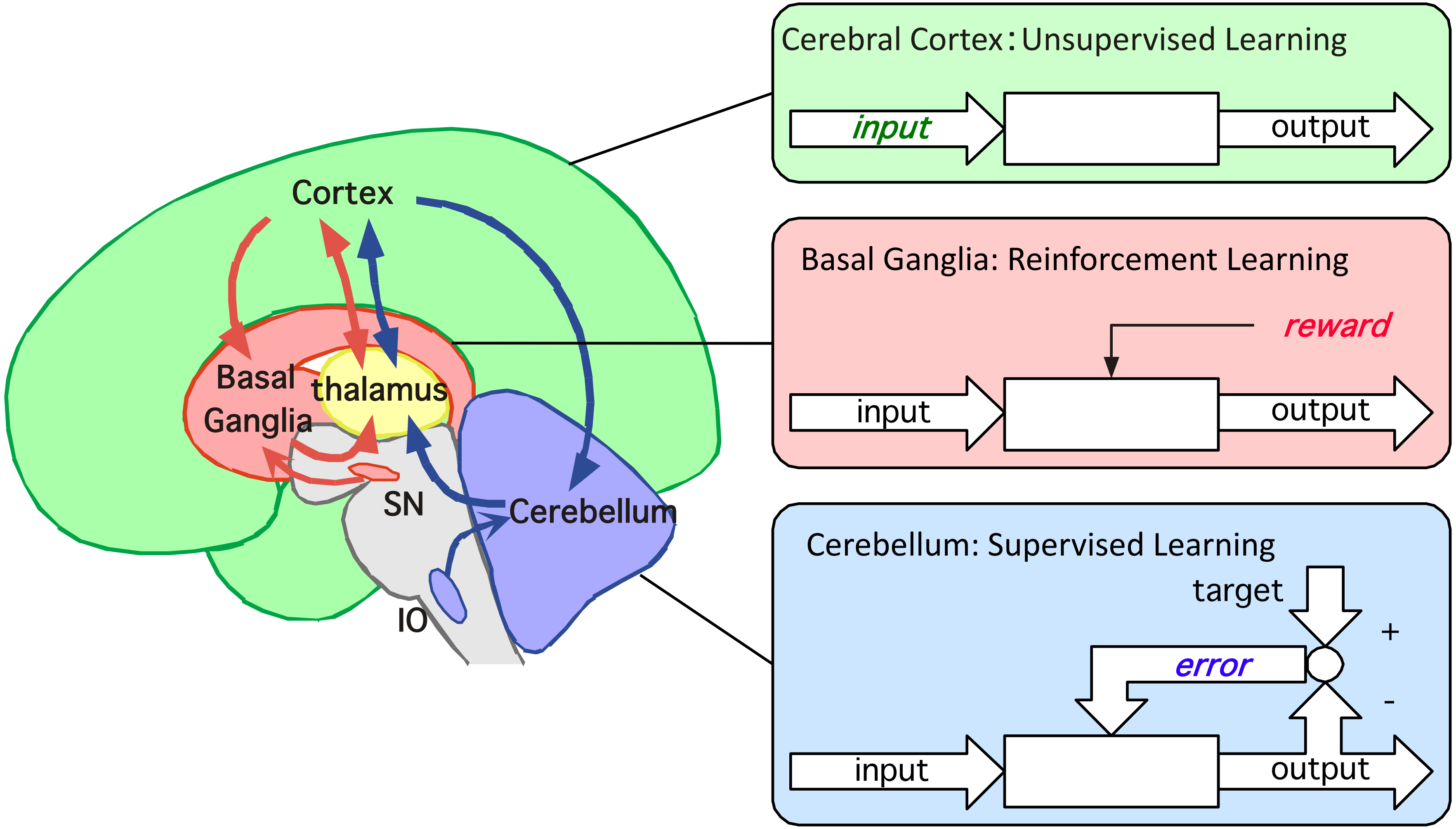




$$Q_{net}(s_A, a_j) = wQ_{MB}(s_A, a_j) + (1 - w)Q_{TD}(s_A, a_j)$$

Specialization by Learning Algorithms

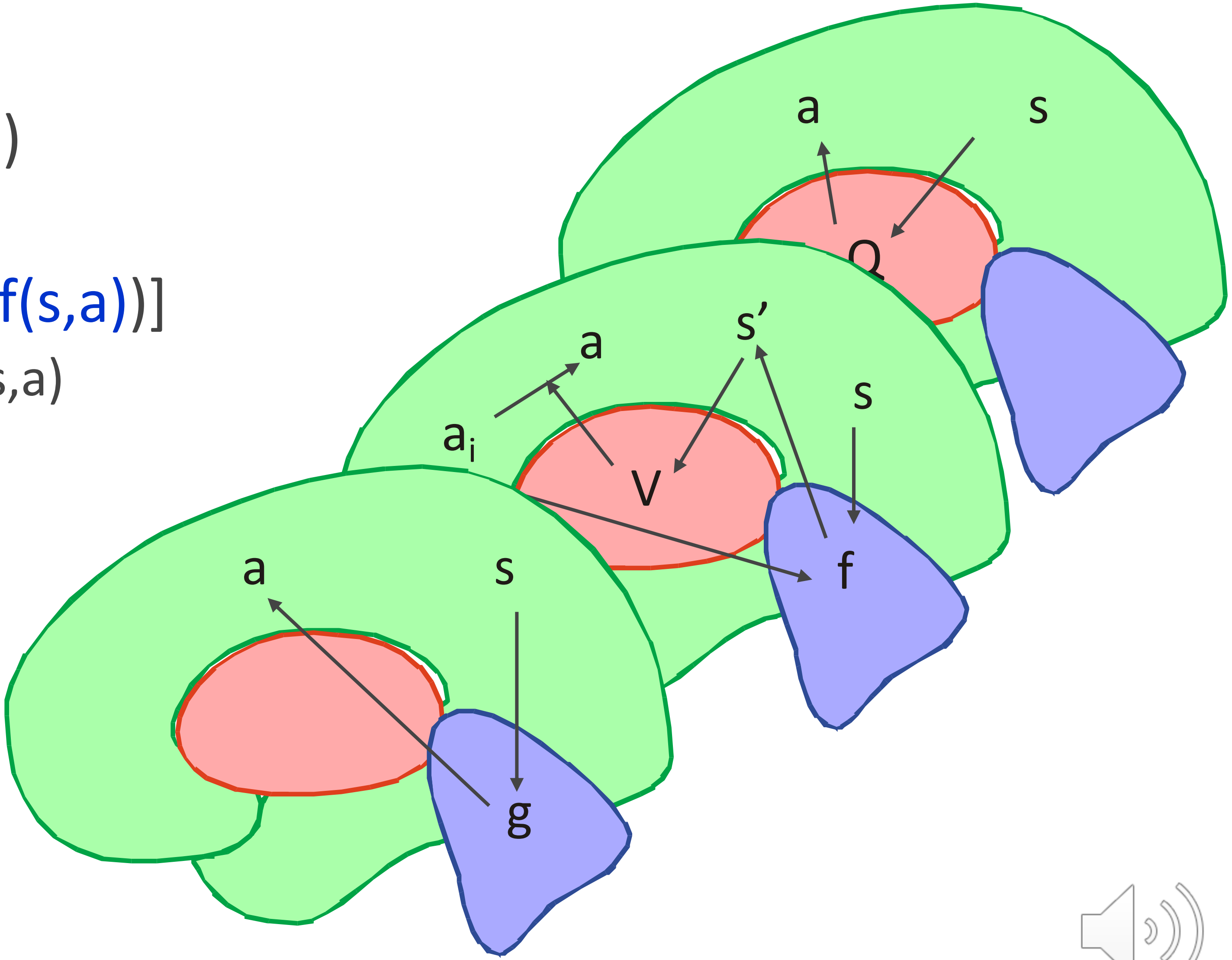
(Doya, 1999)



Multiple Ways of Action Selection

- Model-free
  - $a = \text{argmax}_a\ Q(s,a)$
- Model-based
  - $a = \text{argmax}_a\ [r+V(f(s,a))]$

  forward model: $s'=f(s,a)$
- Memory-based
  - $a = g(s)$



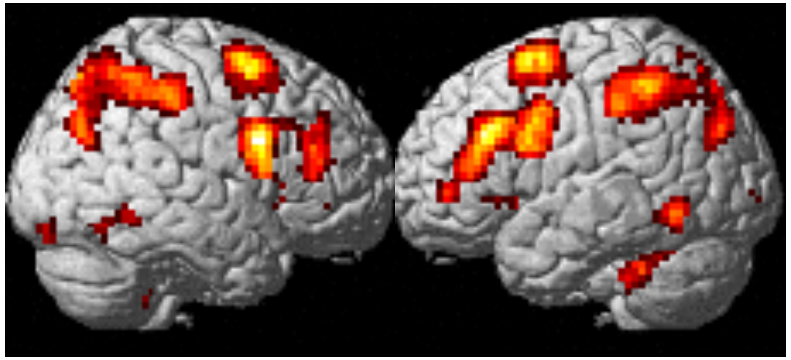
SCIENTIFIC REPORTS

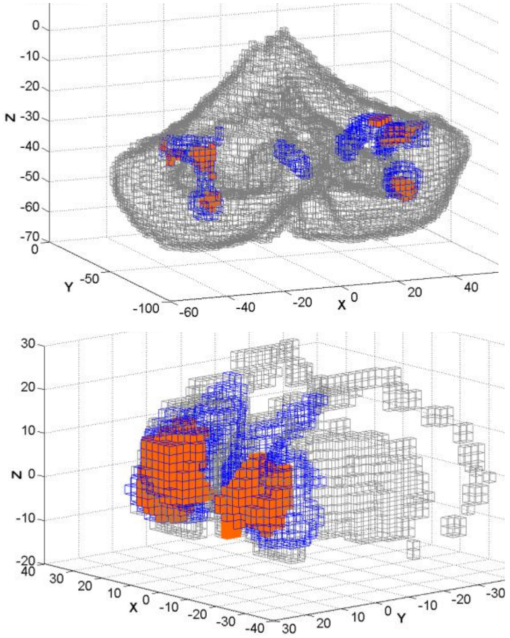
OPEN

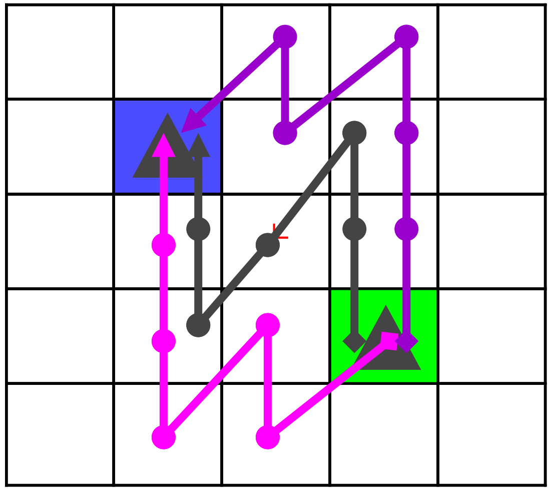
Model-based action planning involves cortico-cerebellar and basal ganglia networks

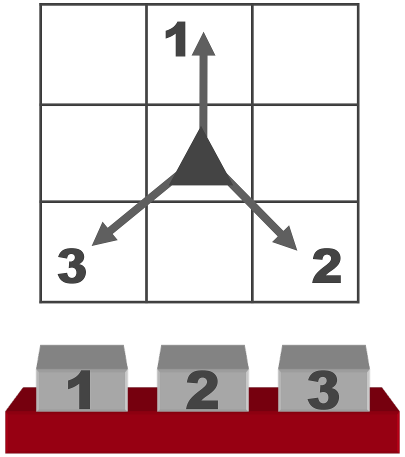
Received: 16 February 2016
Accepted: 19 July 2016

Alan S. R. Fermin[1,2,3], Takehiko Yoshida[1,2], Junichiro Yoshimoto[1,2], Makoto Ito[2], Saori C. Tanaka[4] & Kenji Doya[1,2,3,4]





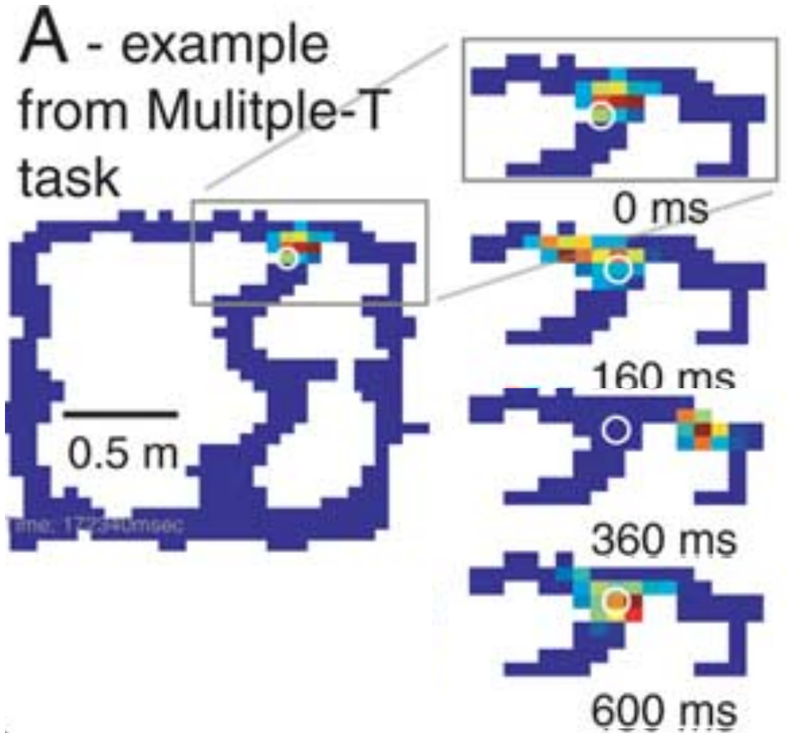




Neuronal Correlates of Mental Simulation

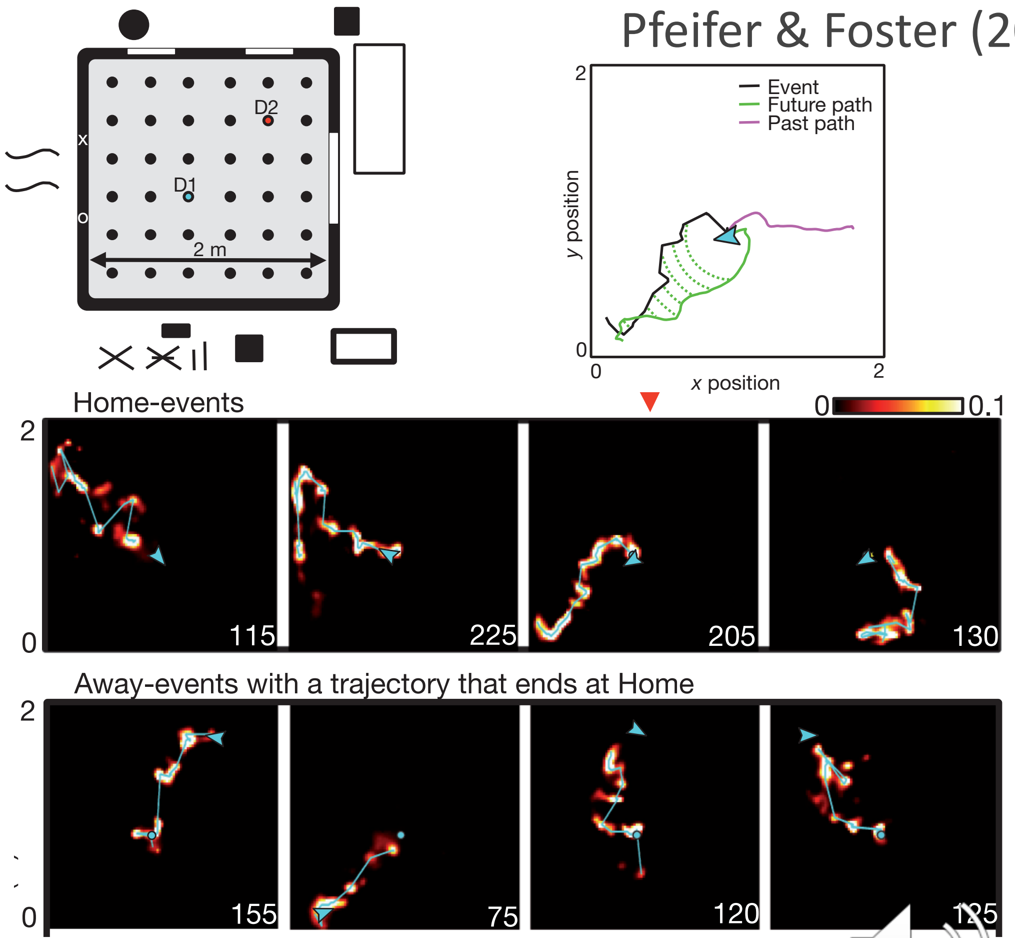
- T-maze

Johnson & Redish (2007)



- Home-Away task

Pfeifer & Foster (2013)

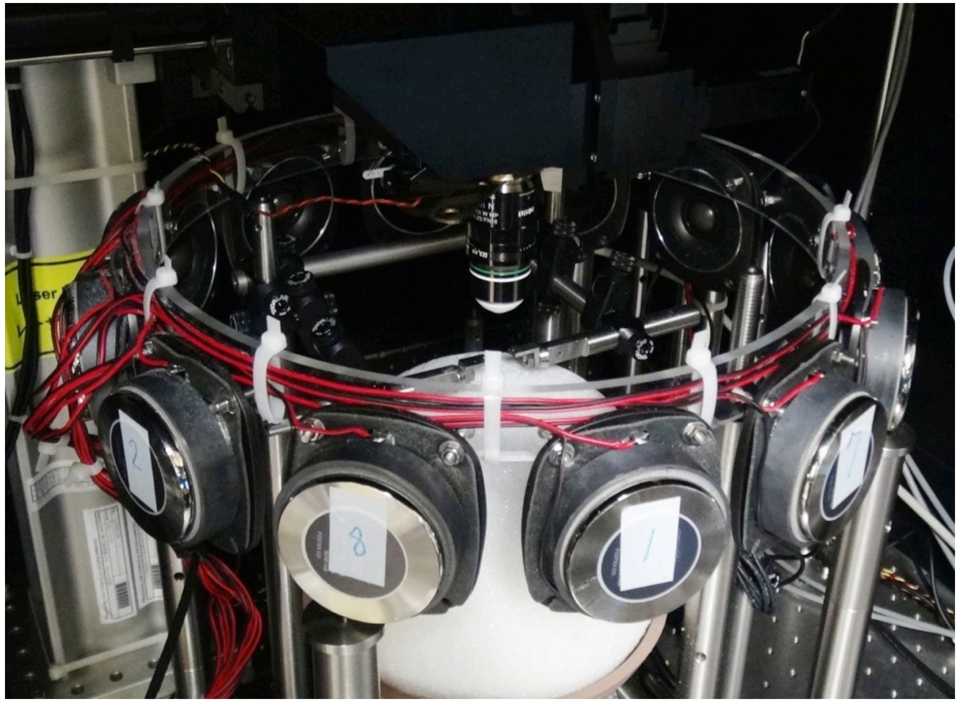


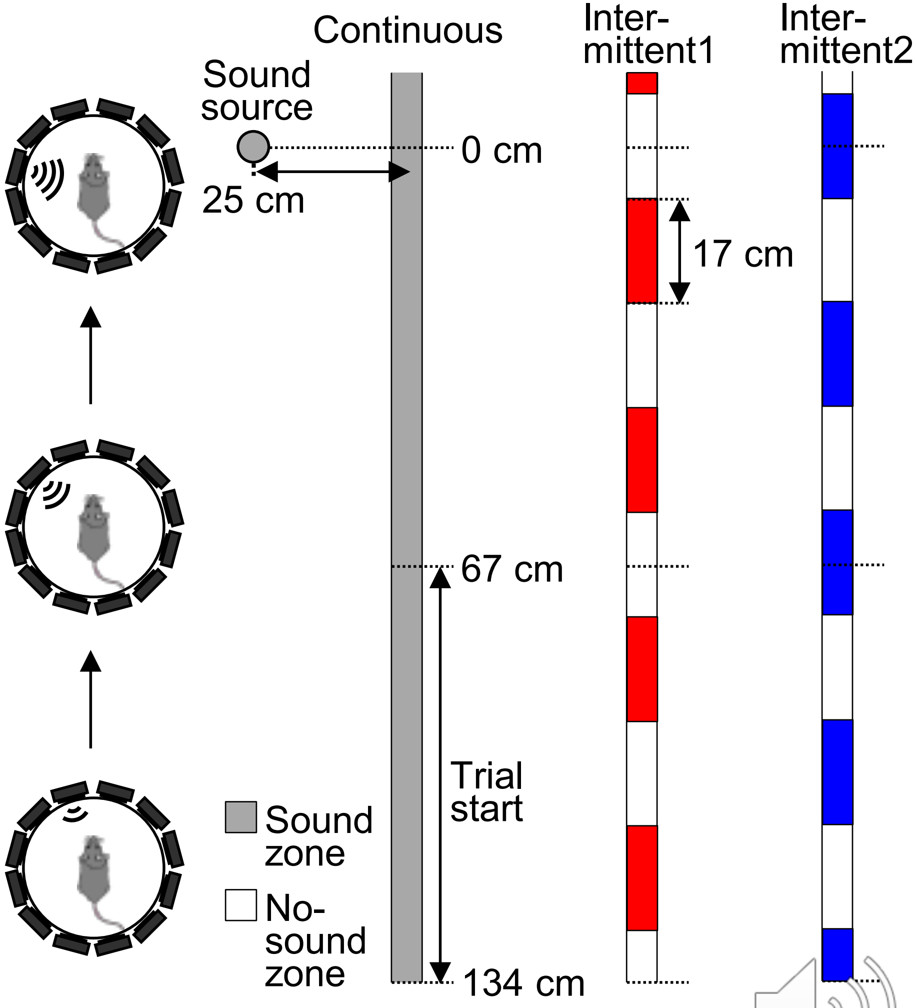
nature neuroscience

Neural substrate of dynamic Bayesian inference in the cerebral cortex

Akihiro Funamizu[1,2], Bernd Kuhn[2] & Kenji Doya[1]

- Auditory virtual environment



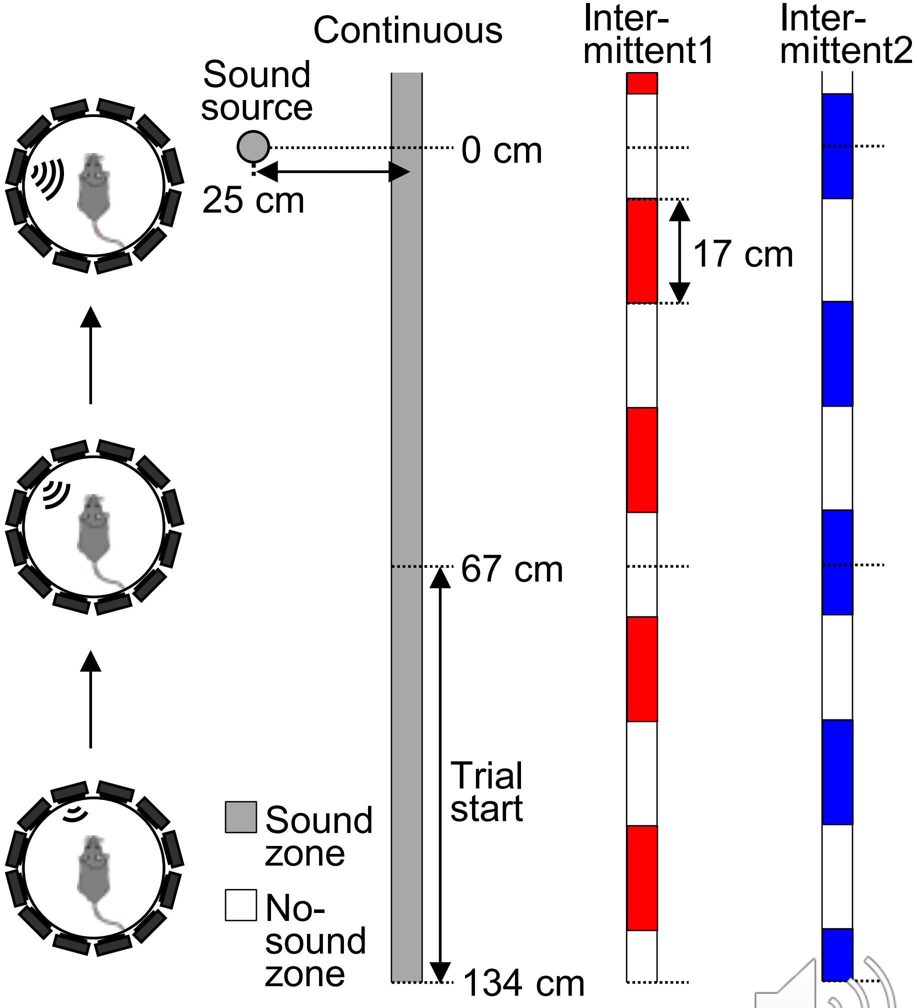


nature neuroscience

Neural substrate of dynamic Bayesian inference in the cerebral cortex

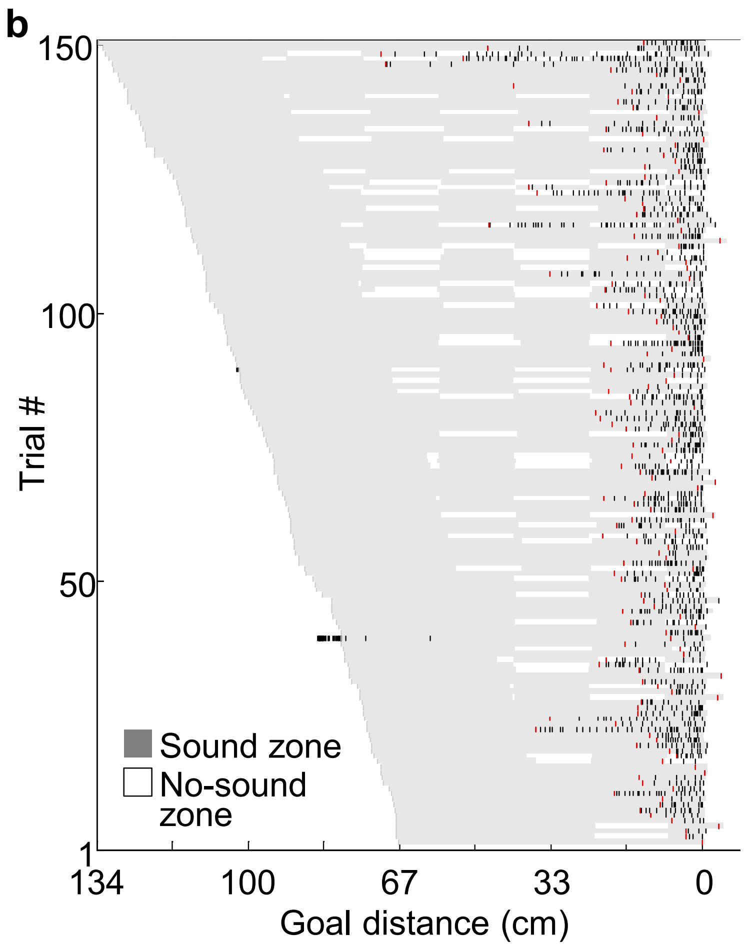
Akihiro Funamizu[1,2], Bernd Kuhn[2] & Kenji Doya[1]

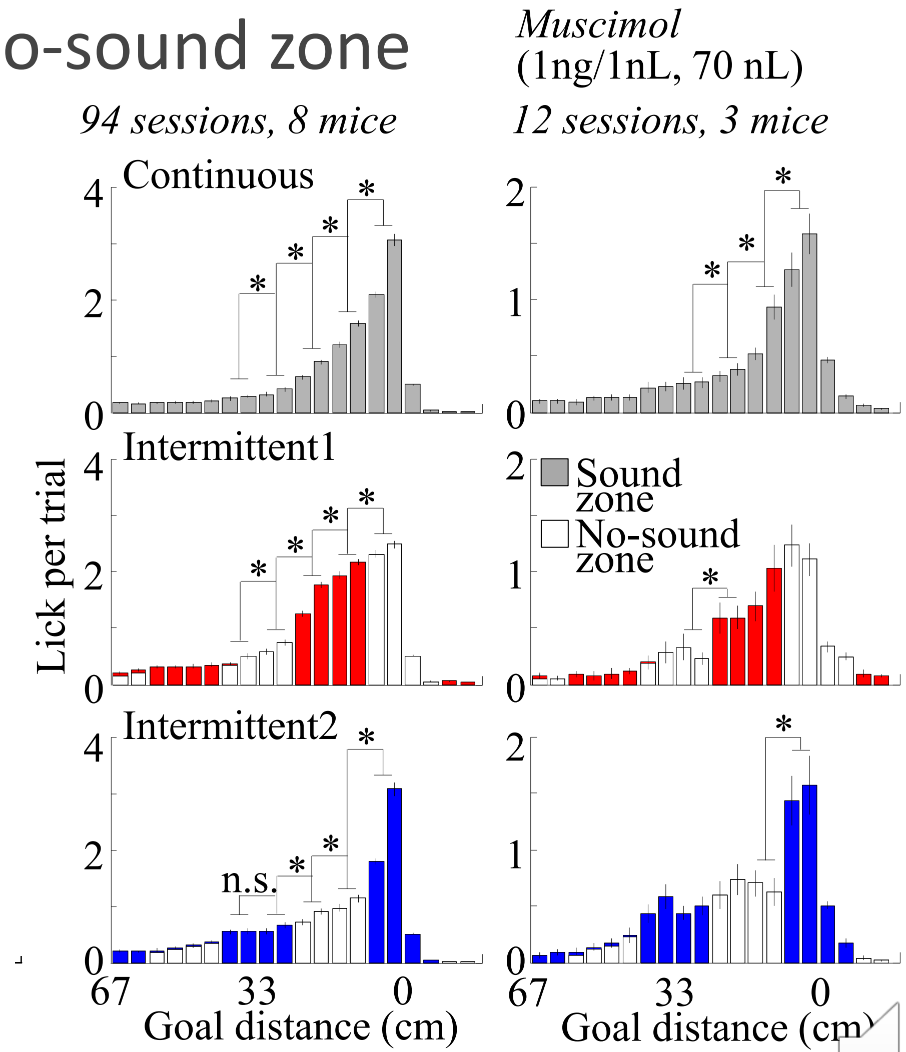
- Auditory virtual environment



Anticipatory Licking

- Mice estimated goal distance in no-sound zone

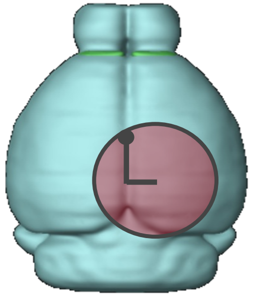


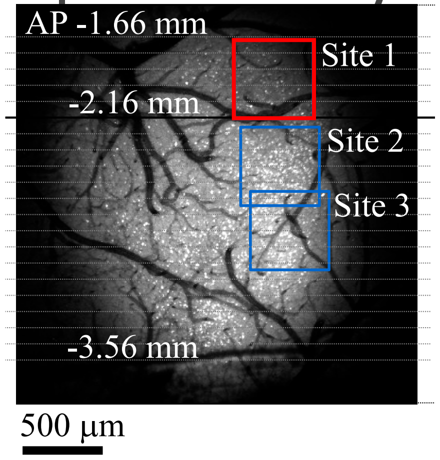


- impaired by muscimol injection in PPC

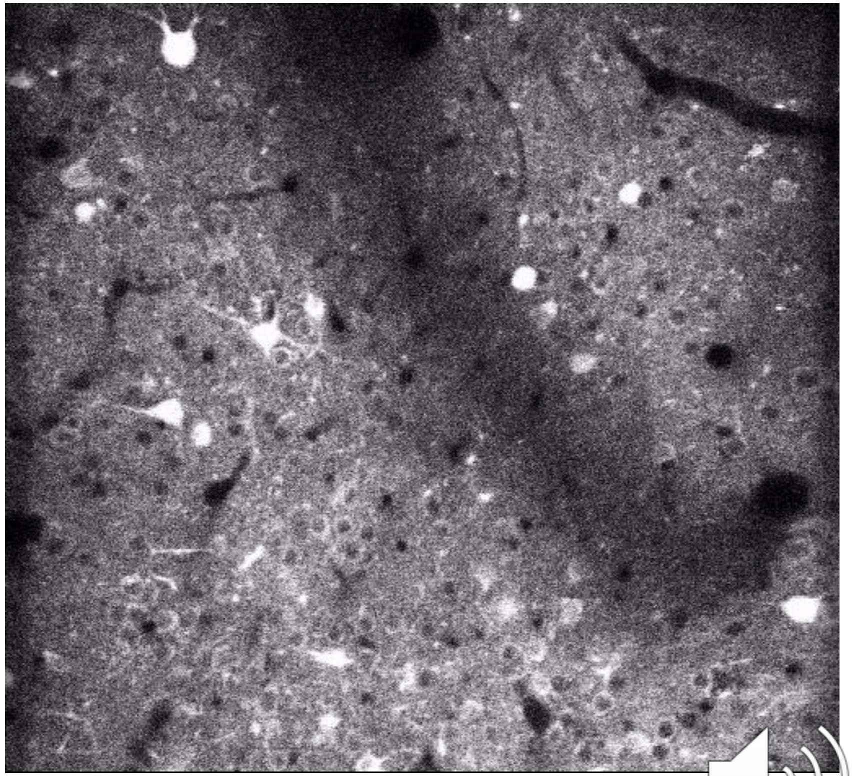
Two-Photon Neural Imaging

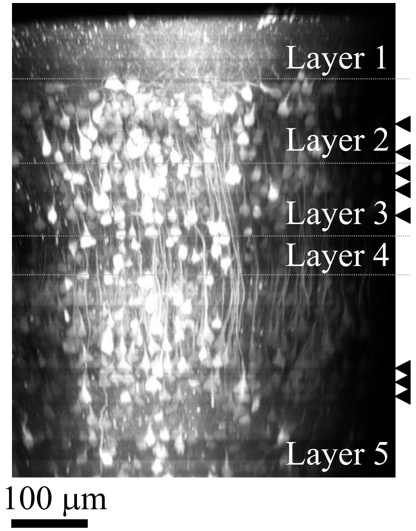
- GCaMP6f expression by AAV





- posterior parietal cortex (PPC)
- auditory-visual cortex (area PM)

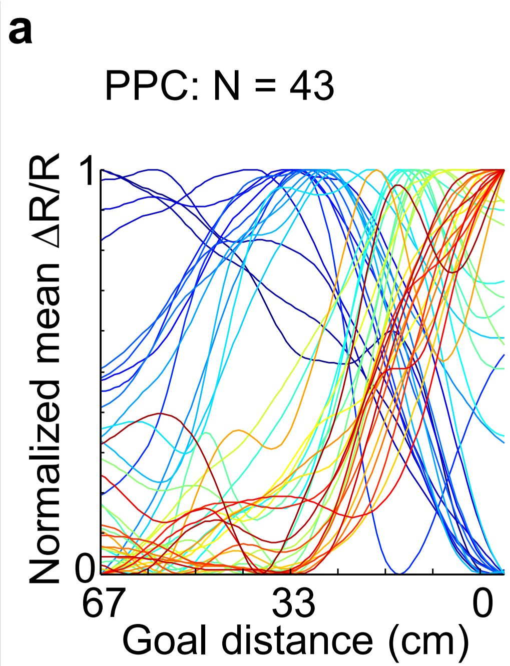


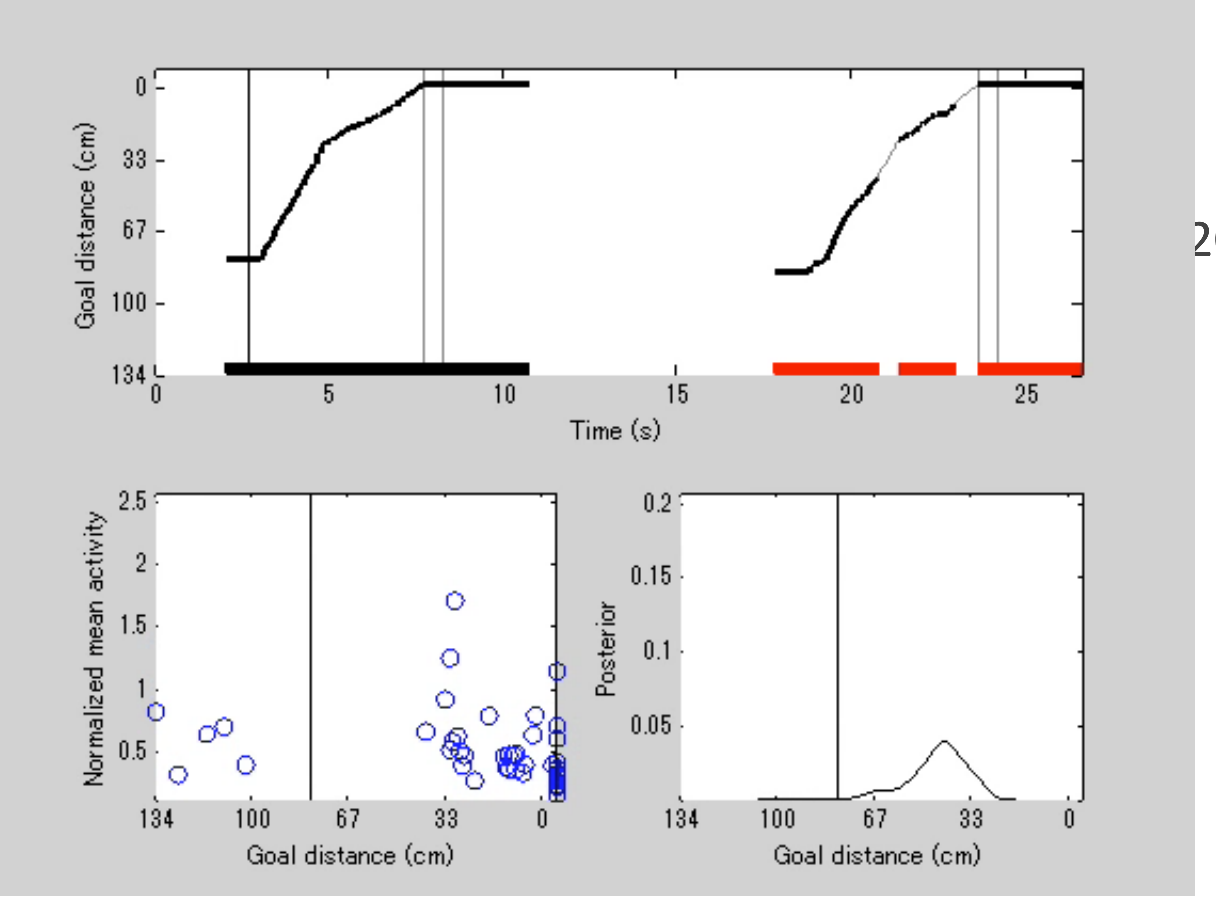


Decoding the Goal Distance

- Neuron i activity f_i at distance x
  - response model $p(f_i|x)$
- Bayesian decoder: $p(x|f_1,$..

2006)





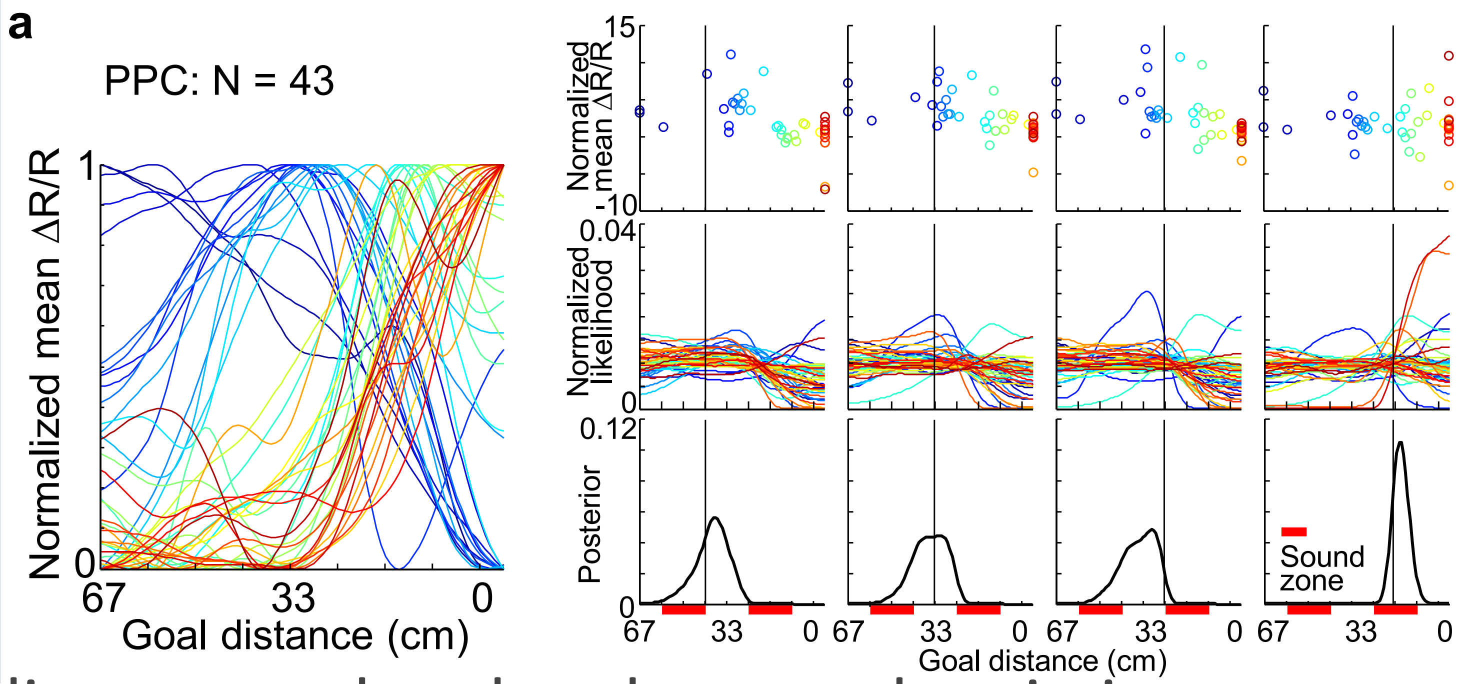
Goal distance (cm)

  - goal distance updated under sound omission

Decoding the Goal Distance

- Neuron *i* activity f_i at distance x
  - response model $p(f_i|x)$
- Bayesian decoder: $p(x|f_1,\ldots,f_N) \propto \prod_i p(f_i|x)p(x)$

(Ma et al., 2006)



  - goal distance updated under sound omission

Two-Photon Imaging: Summary

Auditory virtual navigation task for mice

- estimate goal distance during no-sound phase from its own action using an *internal model*

Two-photon imaging from PPC

- goal distance can be decoded from population activity even during no-sound phase
- variance reduced during sound phase
- characteristic of *dynamic Bayesian inference*

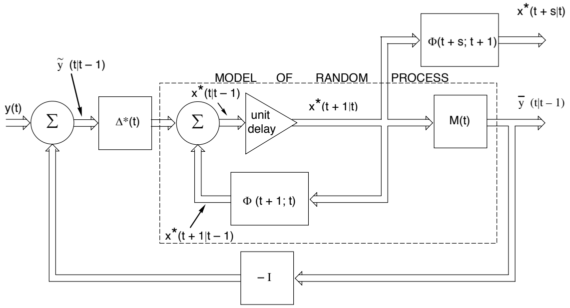
Future

- network mechanisms for action-dependent prediction and sensory-based refinement

Duality of Inference and Control

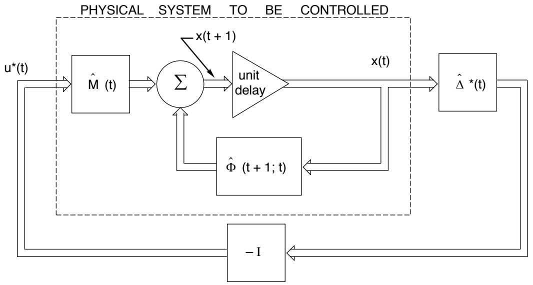
- Optimal filtering (Kalman 1960)

$$\Sigma_{k+1} = S + A\Sigma_k A^\top - A\Sigma_k H^\top \left(P + H\Sigma_k H^\top\right)^{-1} H\Sigma_k A^\top$$

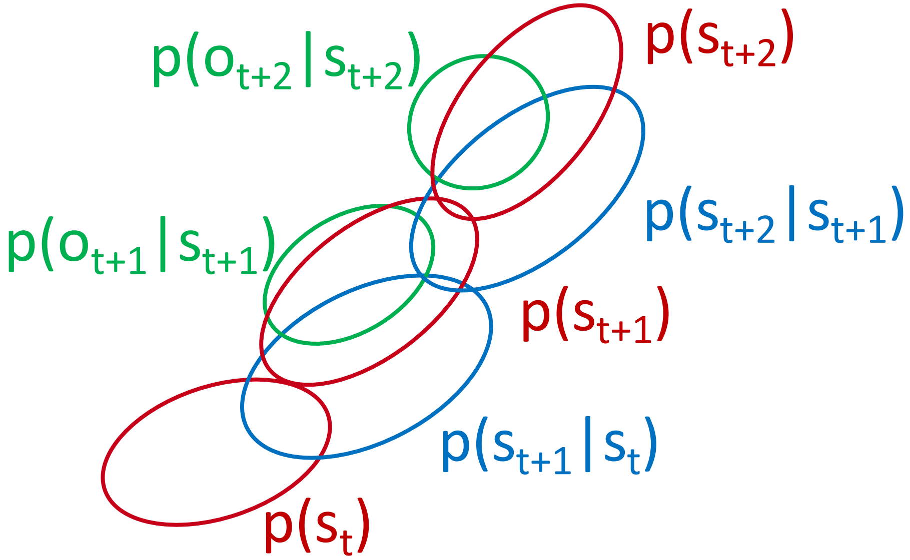


- Optimal control (Bellman et al. 1958)

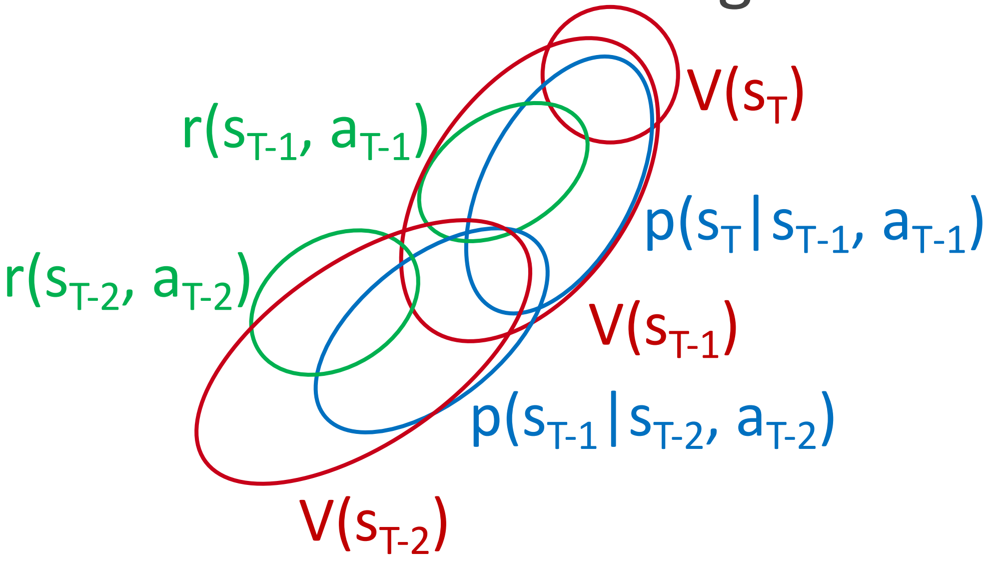
$$V_k = Q + A^\top V_{k+1} A - A^\top V_{k+1} B \left(R + B^\top V_{k+1} B\right)^{-1} B^\top V_{k+1} A$$



- Bayesian inference: log posterior



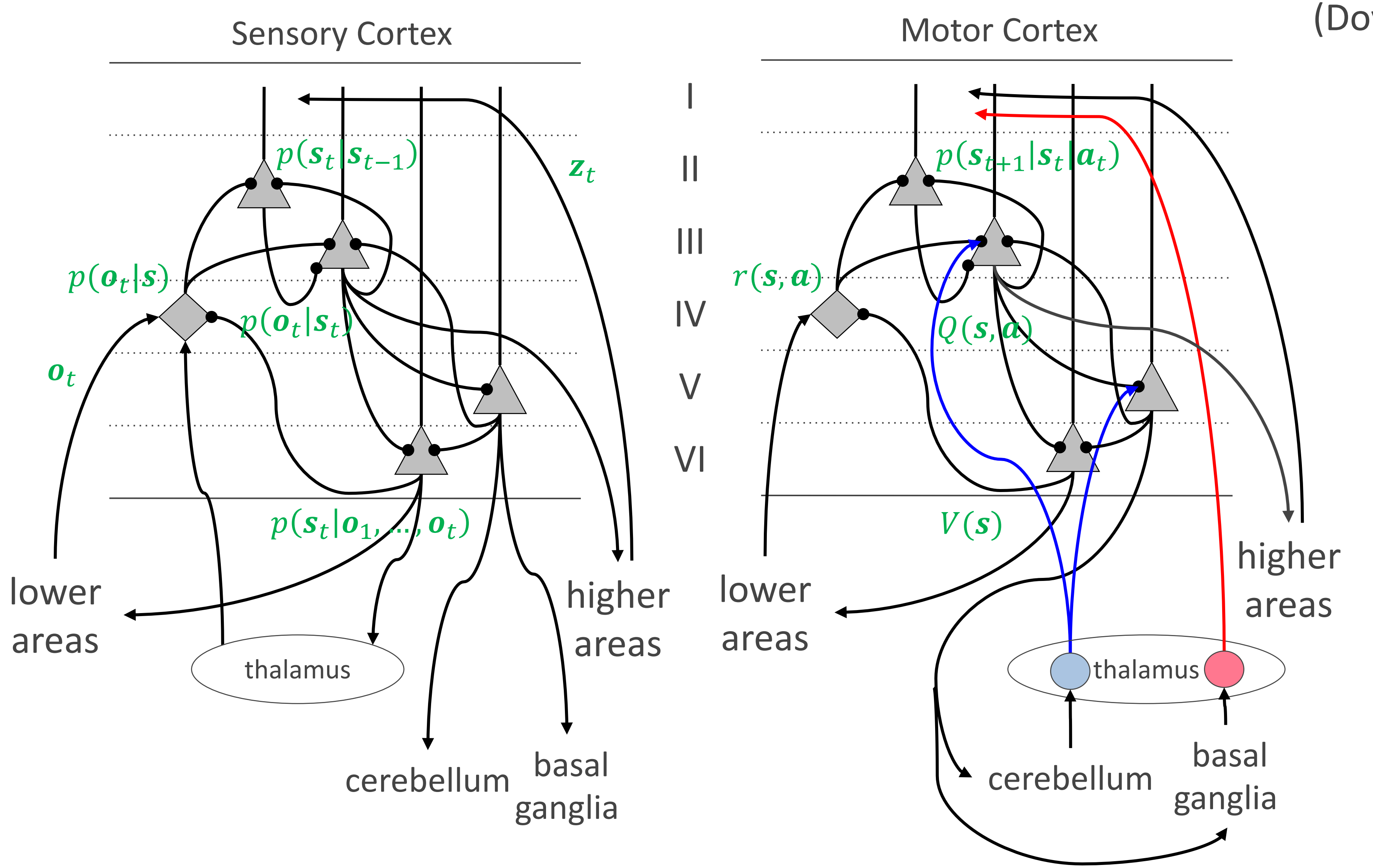
- Reinforcement learning: state value



(Todorov 2007, 08; Toussaint 2009; Levine 2018)

Canonical Cortical Circuits

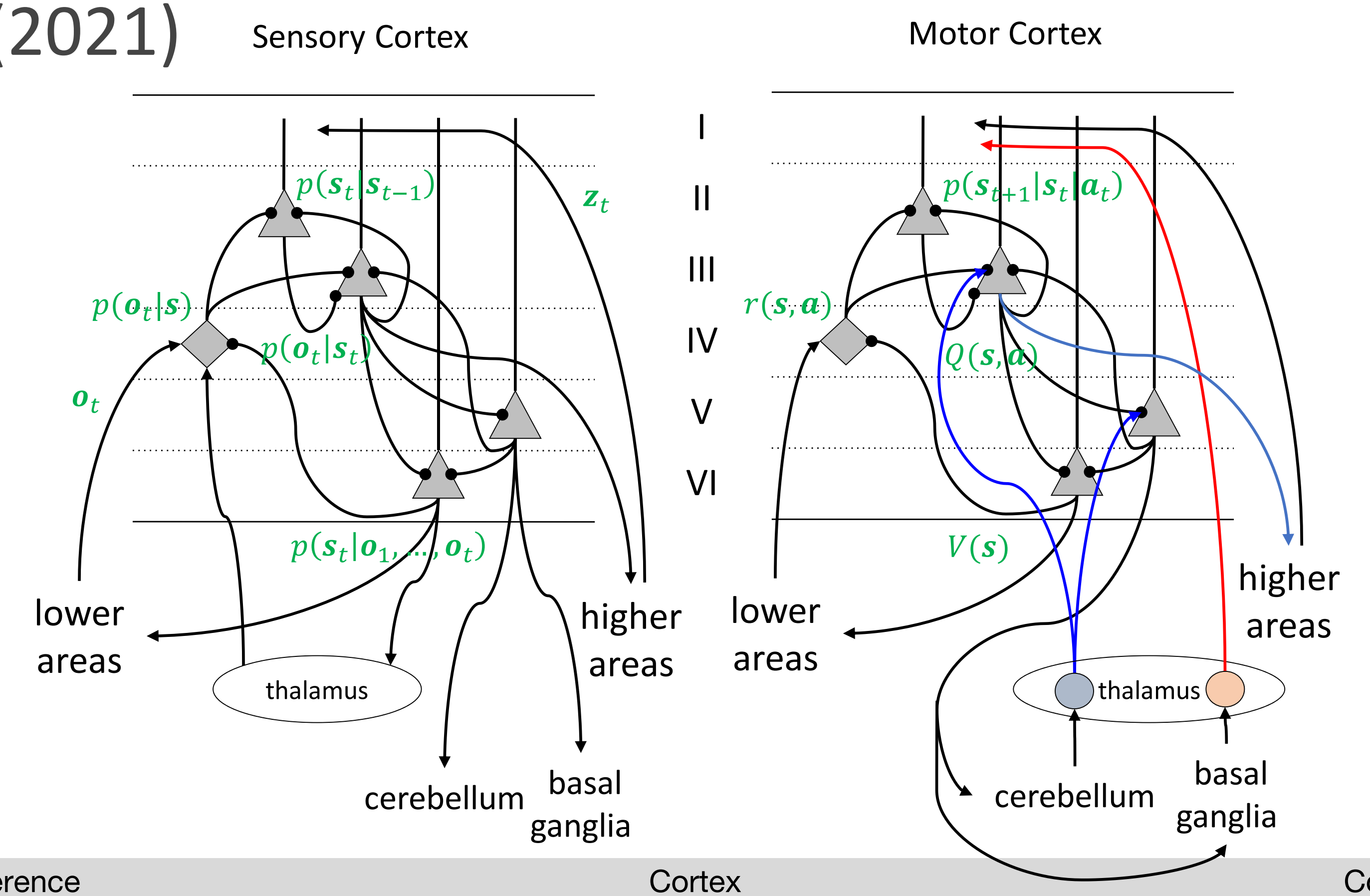
(Doya, 2021)



Canonical cortical circuits and the duality of Bayesian inference and optimal control

Current Opinion in Behavioral Sciences

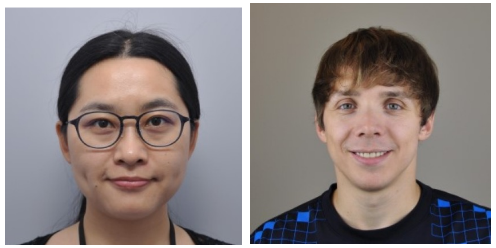
Kenji Doya (2021)

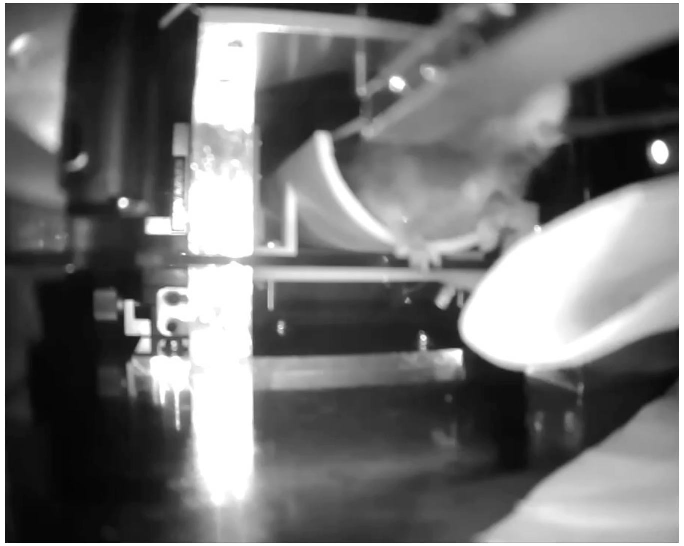


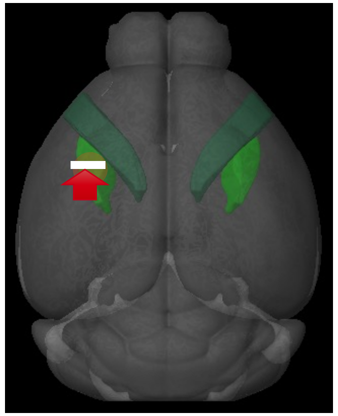
Inference	Cortex	Control
Top-down signal z_t	L1 input	Top-down activation signal
Bottom-up signal $p(o_t \mid s_t)$	L2/3 output	Action value $Q(s,a)$
Predictive model $p(s_t \mid s_{t-1})$	L2/3 connection	Predictive model $p(s_{t+1} \mid s_t, a_t)$
Bottom-up signal o_t	L4 input	Optimality signal O_t
Likelihood $p(o_t \mid s)$	L4 output	Reward function $r(s,a)$
Posterior $p(s_t \mid o_1, \ldots, o_t)$	L5 output	State value $V(s)$
Top-down signal s_t	L6 output	Action $p(a_t \mid s_t)$

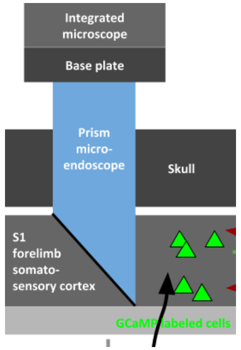
Prism Lens Imaging during Lever Pull Task

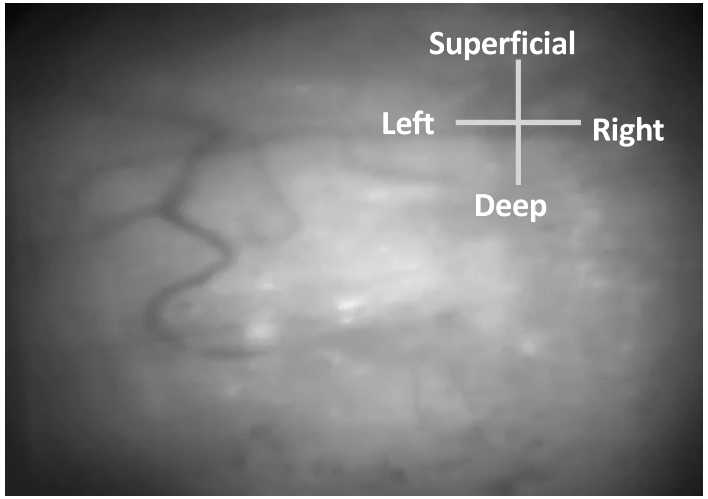
Yuzhe Li, Sergey Zobnin

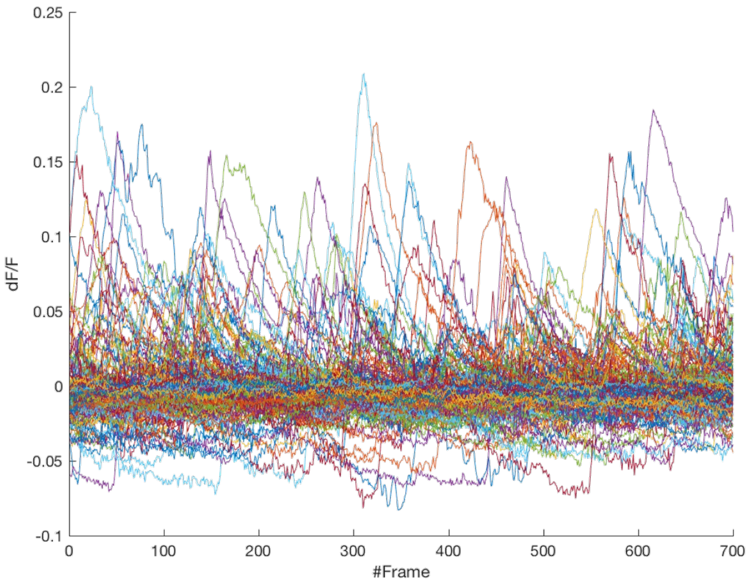


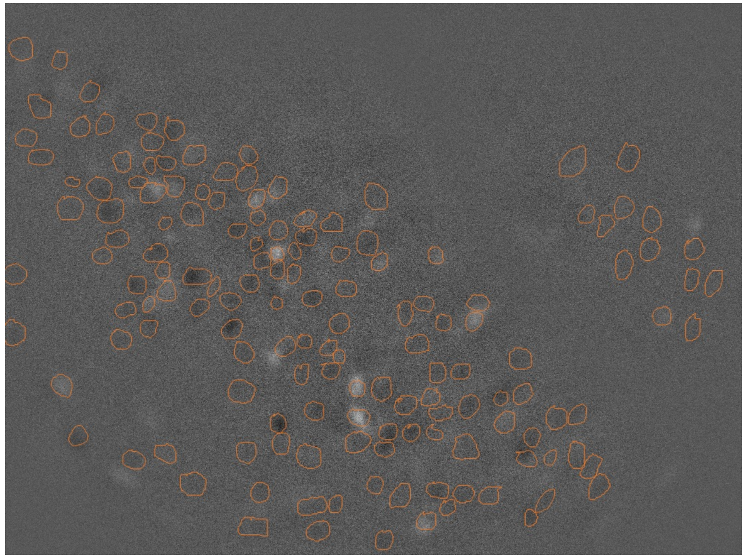






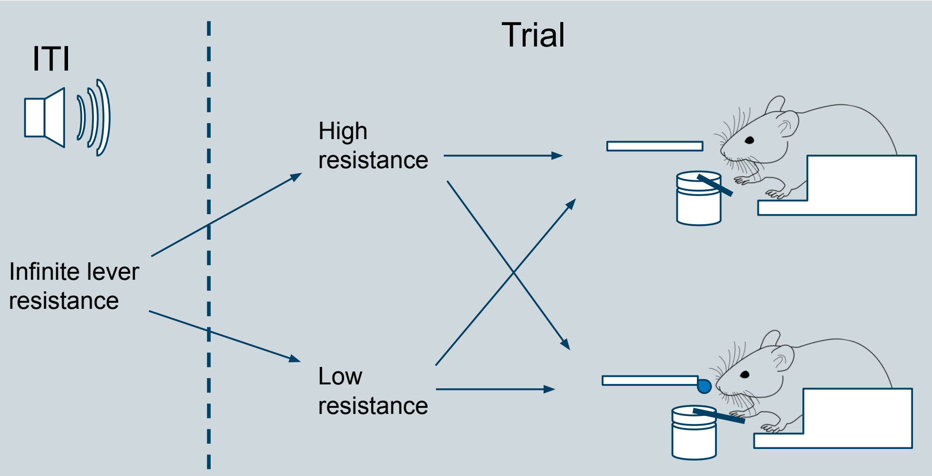




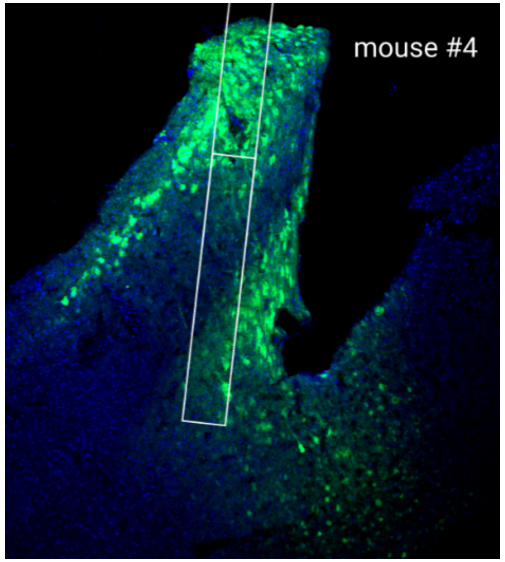


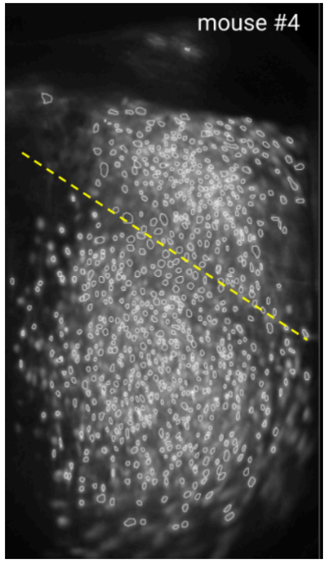
Light/Heavy Lever Pull Task

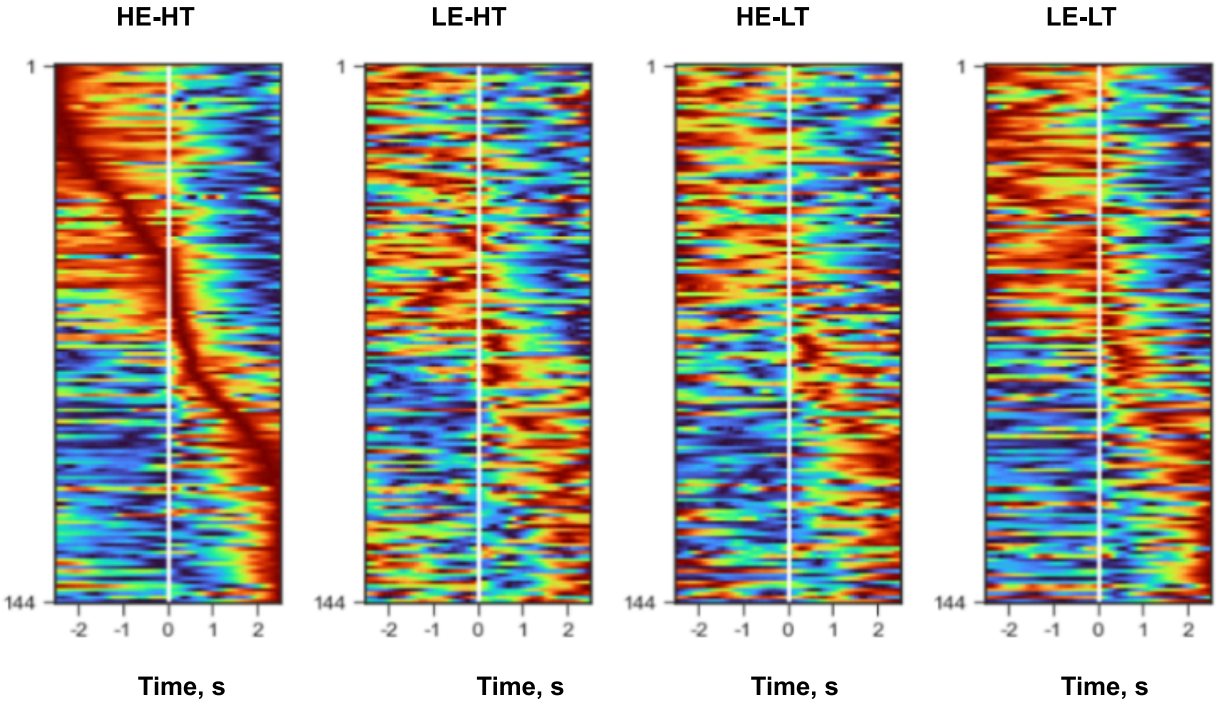
Sergey Zobnin



	Trials: H - heavy resistance, L - light resistance; standard, odd
Light session:	L L L L L L H L L L L L L L L L H L L L L L L L L L L L L L L H L L
Heavy session:	H H H L H H H H H H H H H H H L H H H H H H H L H H L H H
Uniform session:	H H L H L L L H L L H H L H H H H L L H L L L H L H H L H L H
Roving oddball paradigm (ROP):	H H H H H H H H H H L L L L L L L H H H H H H H H L L L L L L Heavy train, Light train



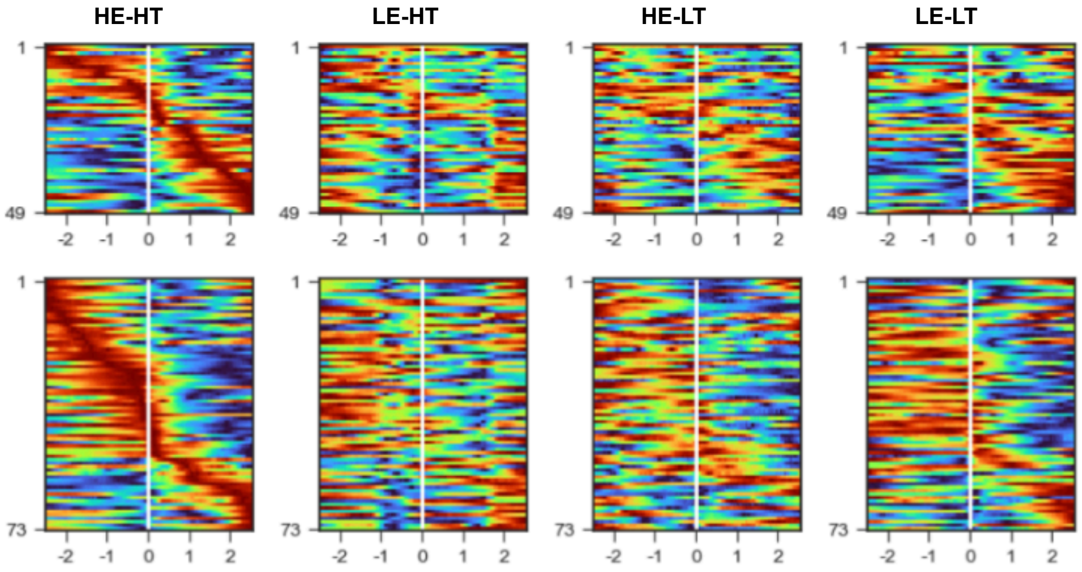


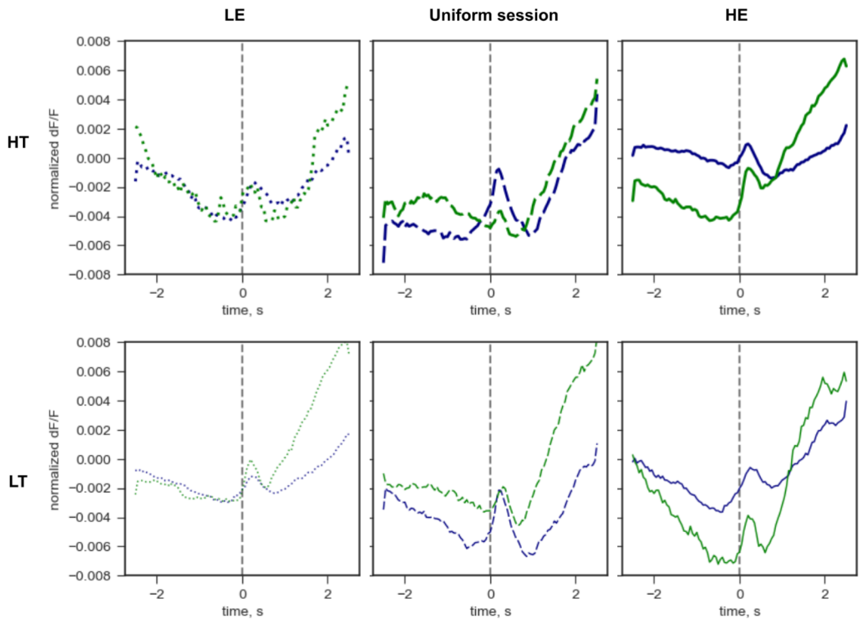


Expected and Actual Trial Type Coding

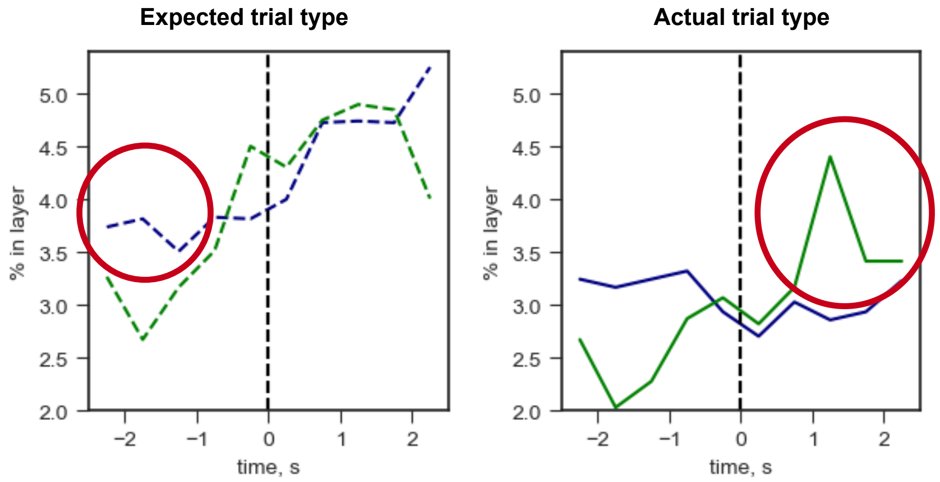
superficial

deep





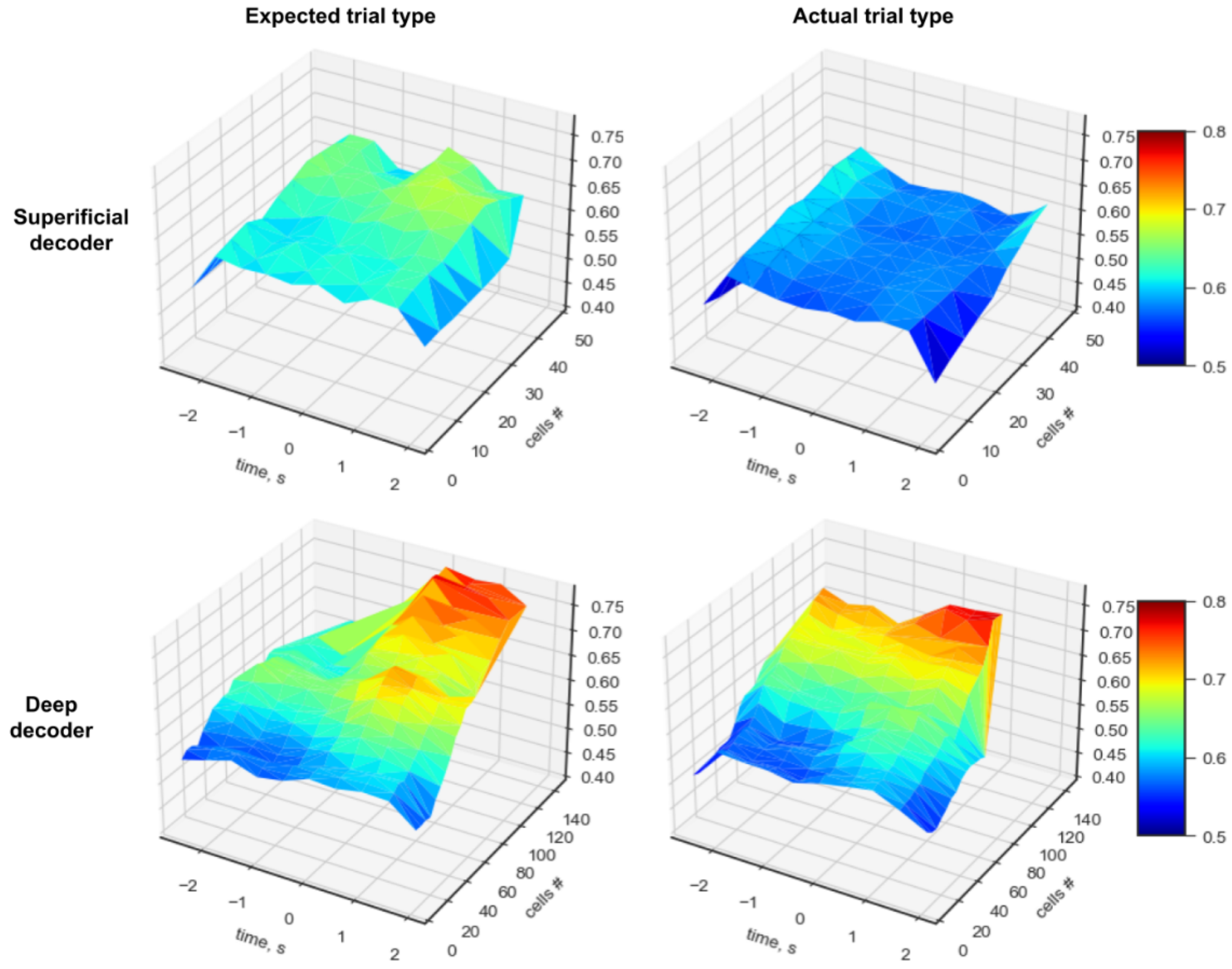
Encoding analysis



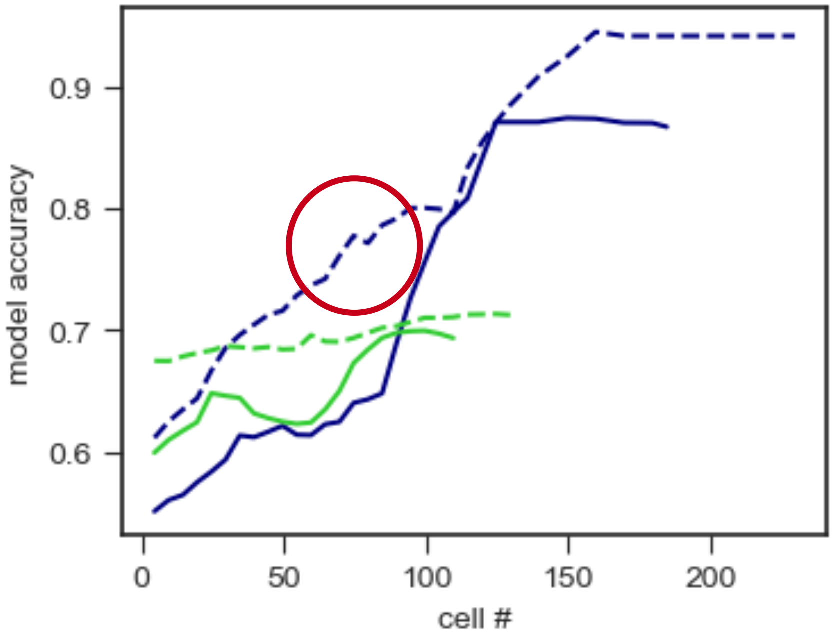
- More deep neurons code expected trial type before action
- More superficial neurons code actual trial type after action

Population Decoding

At different time points



Peak amplitude after pull



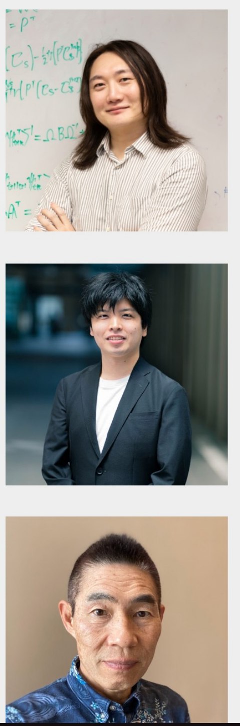
- Better decoding of expected trial type from deep neurons

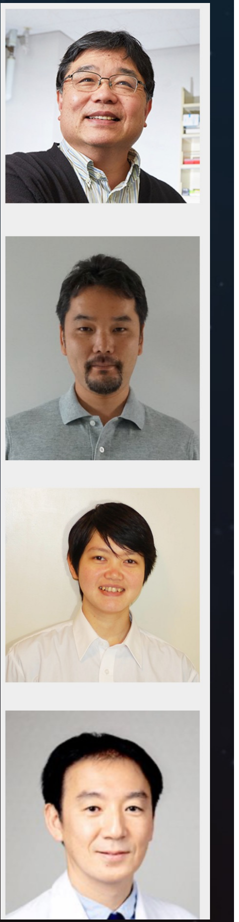
Home About this project News Research Teams Invited Research Teams Publication Contact Japanese

unifiedtheory.jp

Development and validation of a unified theory of prediction and action

Transformative Research Area (A)：unified theory of prediction and action





Reinforcement Learning

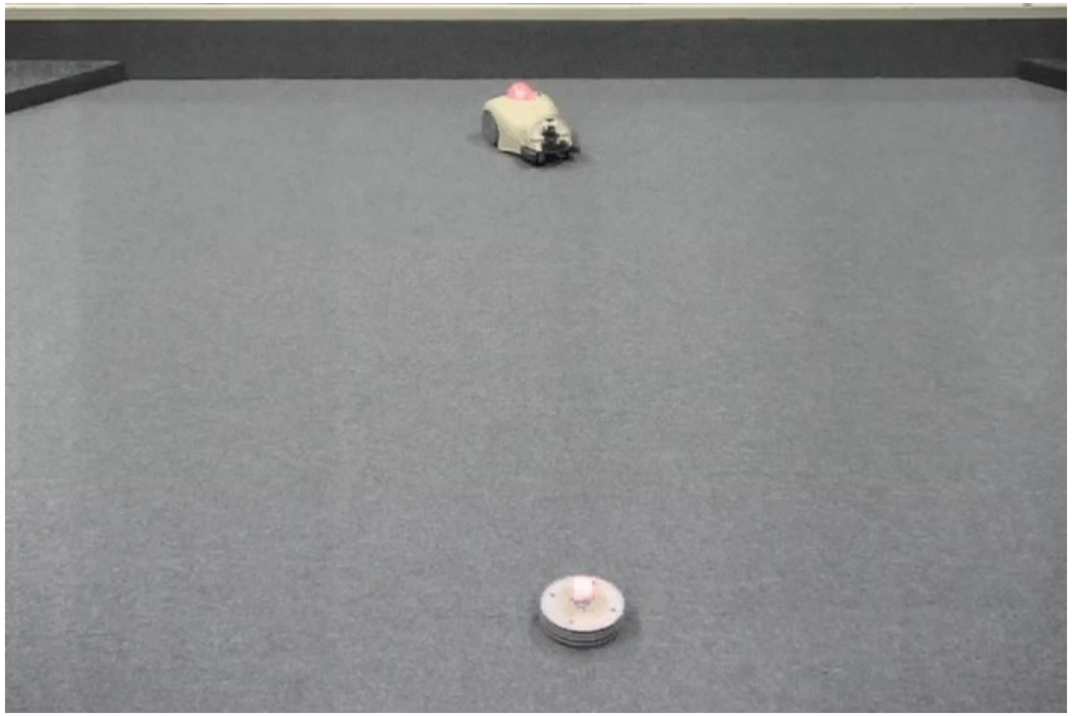
- Predict reward: *value function*
  - $V(s) = E[\ r(t) + \gamma r(t+1) + \gamma^2 r(t+2)\ldots|\ s(t)=s]$
  - $Q(s,a) = E[\ r(t) + \gamma r(t+1) + \gamma^2 r(t+2)\ldots|\ s(t)=s,\ a(t)=a]$
- Select action
  - *greedy*: $a = \operatorname{argmax} Q(s,a)$
  - *Boltzmann*: $P(a|s) \propto \exp[\ \beta\ Q(s,a)]$
- Update prediction: *temporal difference* (*TD*) *error*
  - $\delta(t) = r(t) + \gamma V(s(t+1)) - V(s(t))$
  - $\Delta V(s(t)) = \alpha\ \delta(t)$
  - $\Delta Q(s(t),a(t)) = \alpha\ \delta(t)$

How to implement these steps?

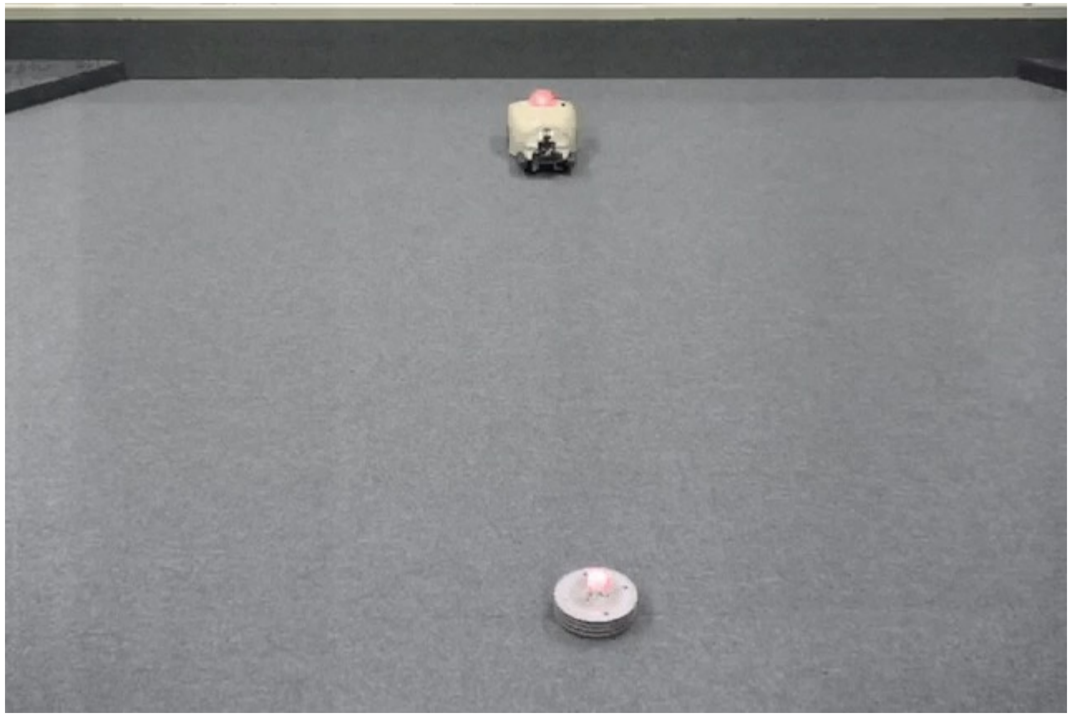
How to tune these parameters?

Temporal Discount Factor γ

- Large γ
  - reach for far reward

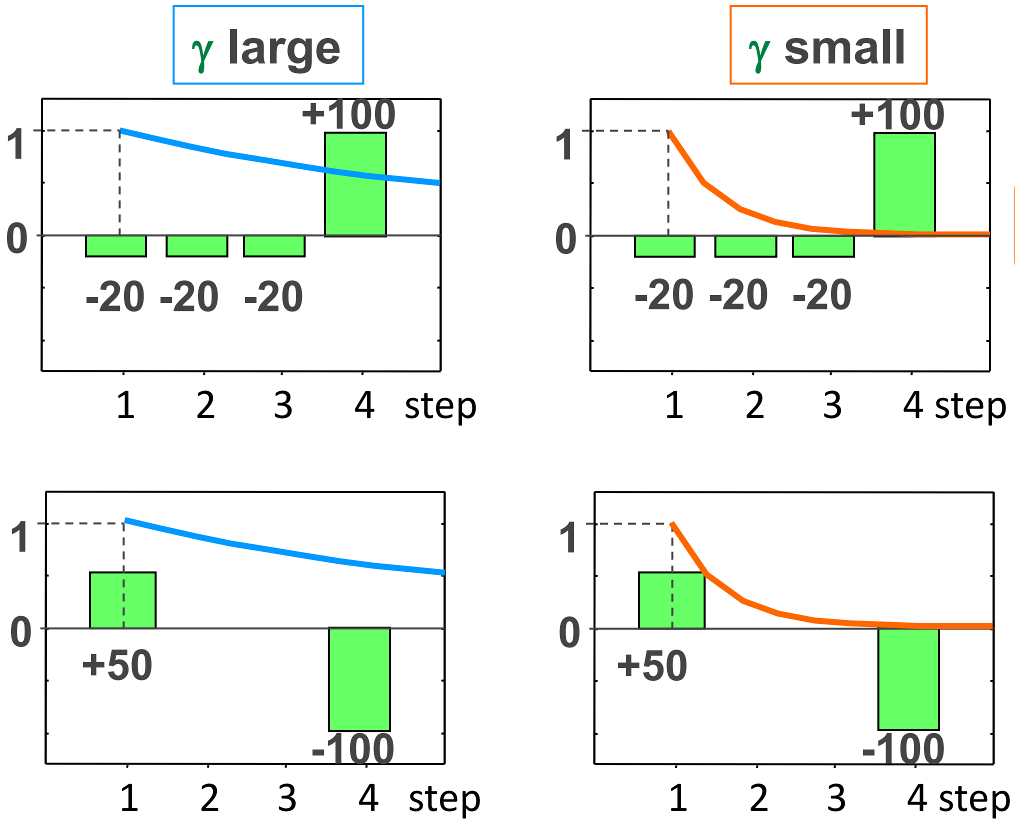


- Small γ
  - only to near reward



Temporal Discount Factor γ

- $V(t) = E[\, r(t) + \gamma r(t+1) + \gamma^2 r(t+2) + \gamma^3 r(t+3) + \ldots]$
  - controls the 'character' of an agent



Depression?

no pain, no gain! — V =18.7

better stay idle — V =-25.1

Impulsivity?

stay away from danger — V = -22.9

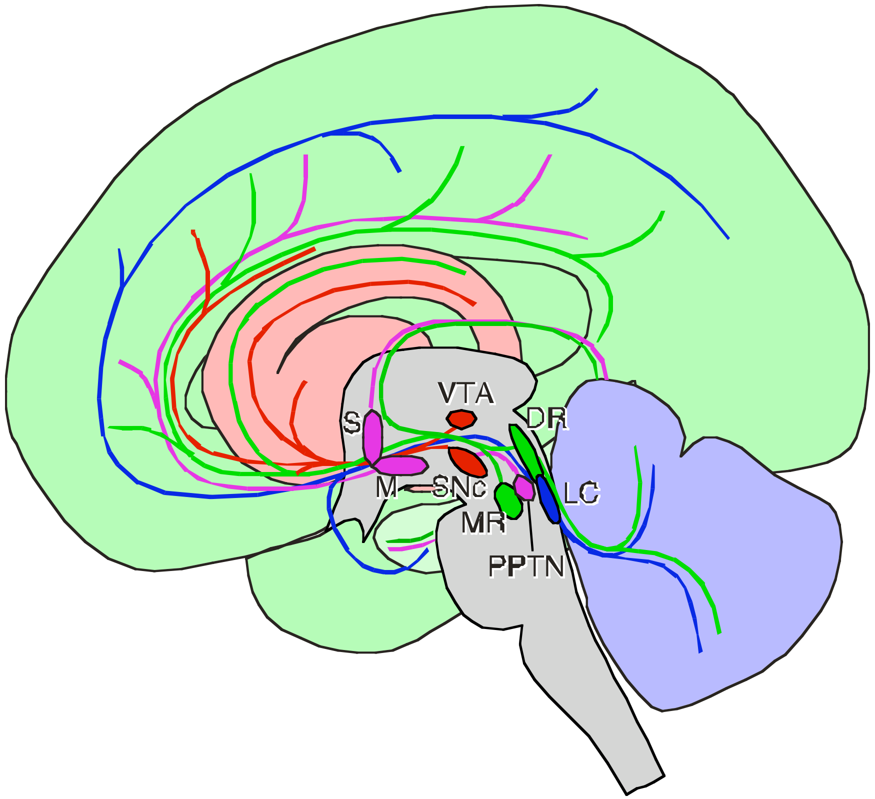
can't resist temptation — V = 47.3

Serotonin?

Neuromodulators for Metalearning

(Doya, 2002)

- *Metaparameter* tuning is critical in RL
  - How does the brain tune them?



Dopamine: TD error δ

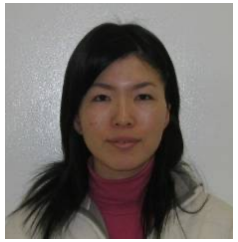
Acetylcholine: learning rate α

Noradrenaline: exploration β

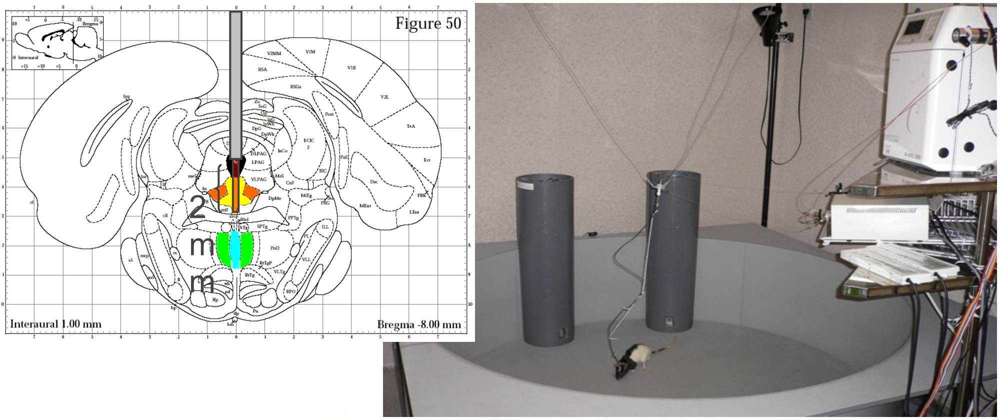
Serotonin: temporal discount γ

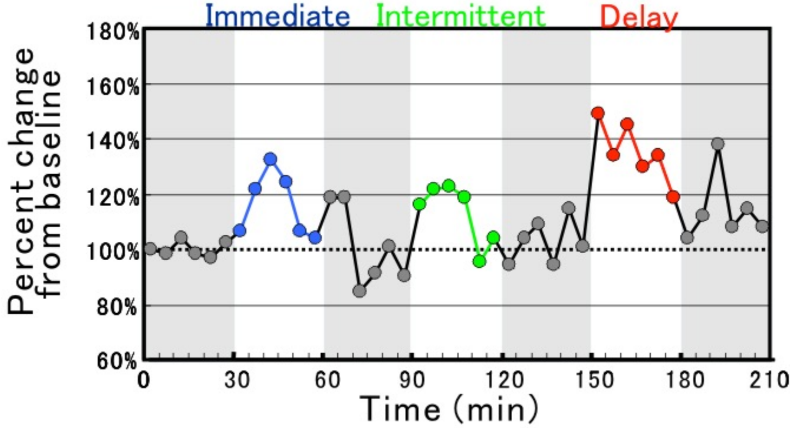
Chemical Measurement/Control

(Kayoko Miyazaki et al., 2011, 2012)



Microdialysis measurement

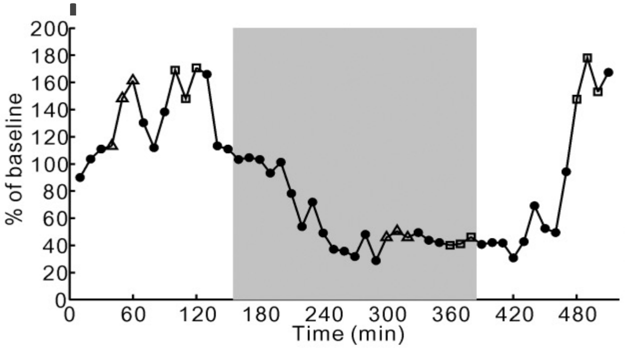


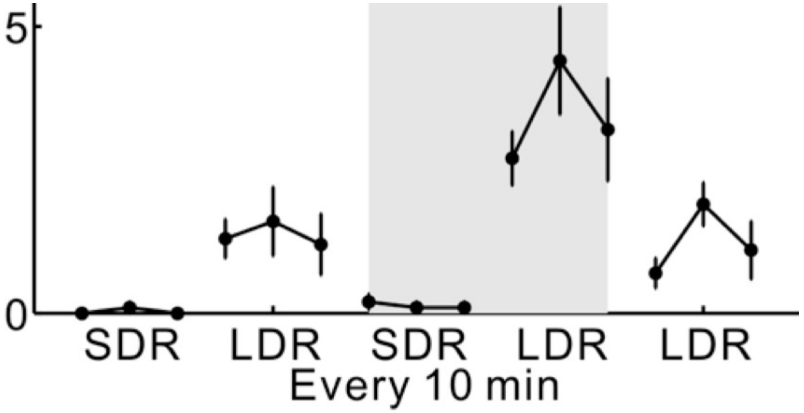


- Serotonin release increased in delayed reward task

Serotonin neuron blockade

- 5HT1A agonist in dorsal r



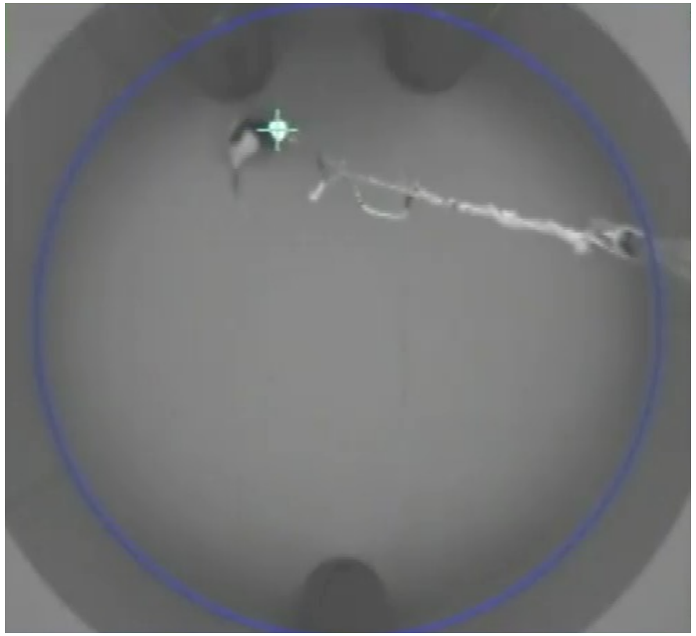


- Waiting error increased in long-delayed reward trials

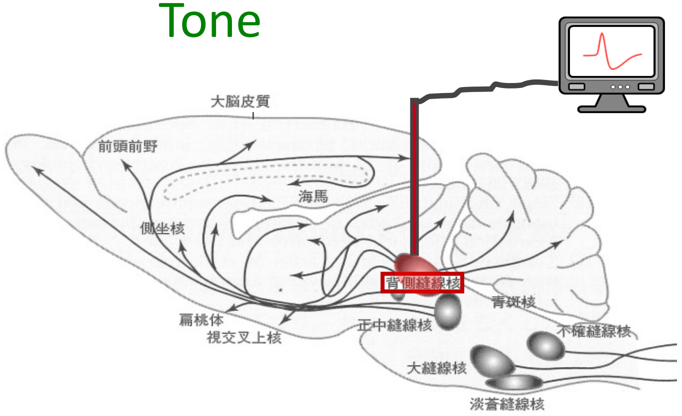
Dorsal Raphe Neuron Recording

(Miyazaki et al. 2011 JNS)

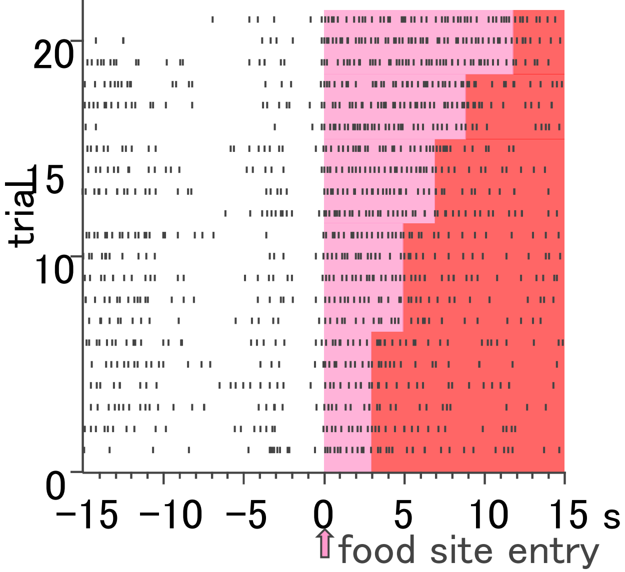
Food Water



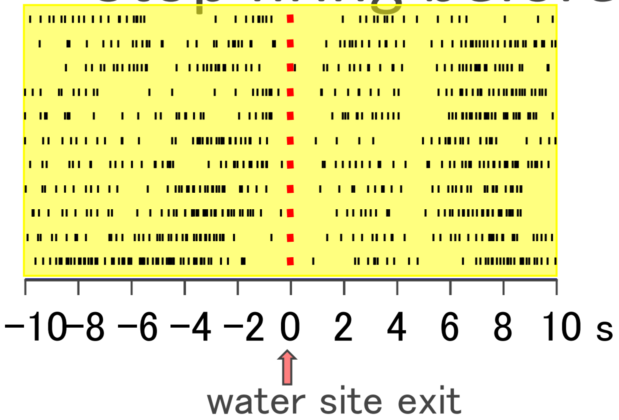
Tone



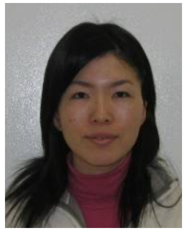
- Keep firing while waiting



- Stop firing before giving up

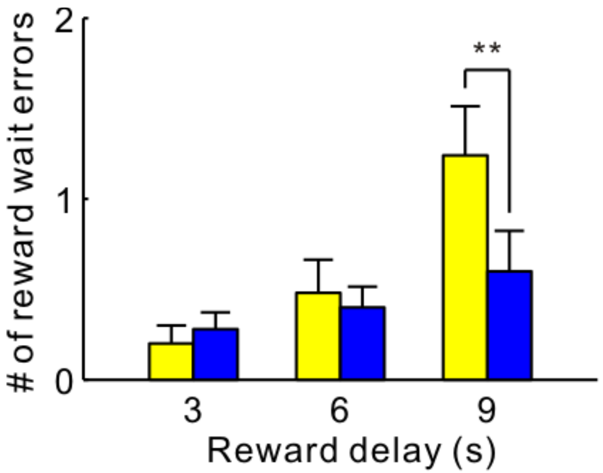


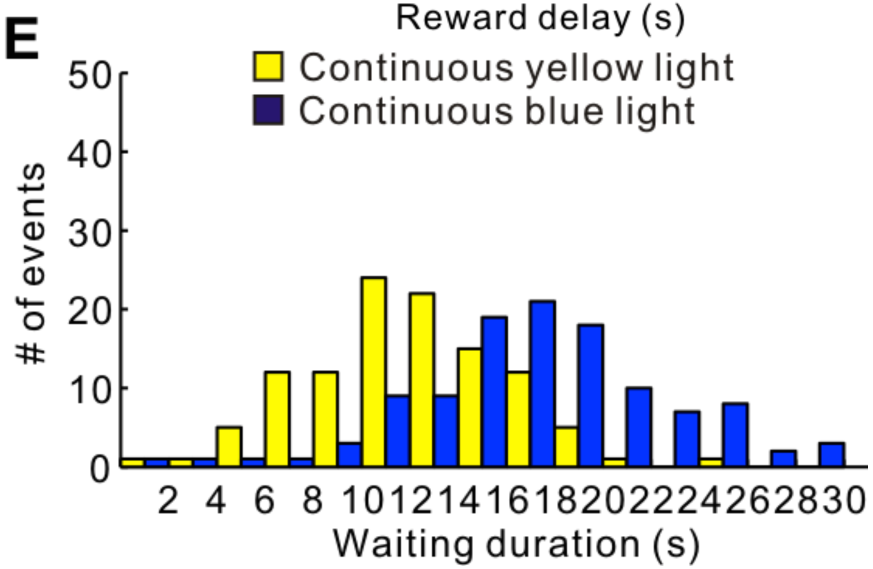
Optogenetic Stimulation of Serotonin Neurons



(Miyazaki et al., 2014, Current Biology)

- Reward Delay Task (3, 6, 9, ∞ sec)





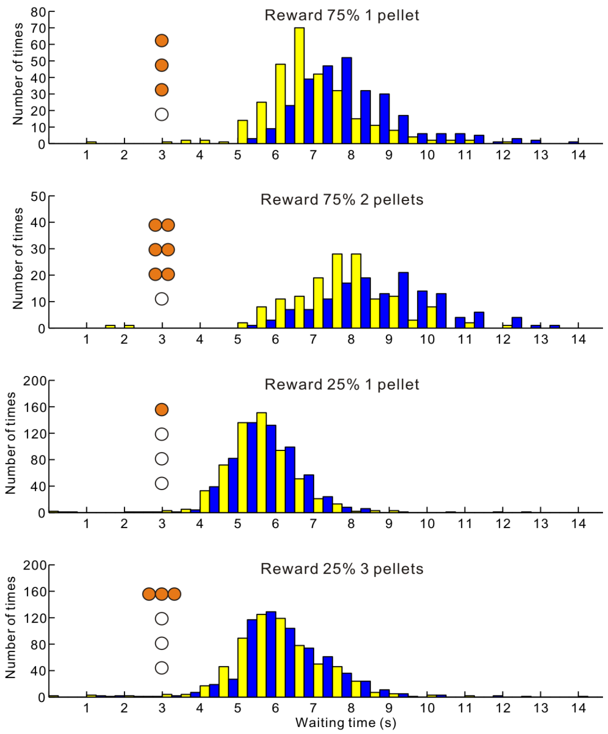
- 3 sec: success
- omission: 12.1 s
- omission: 20.8 s

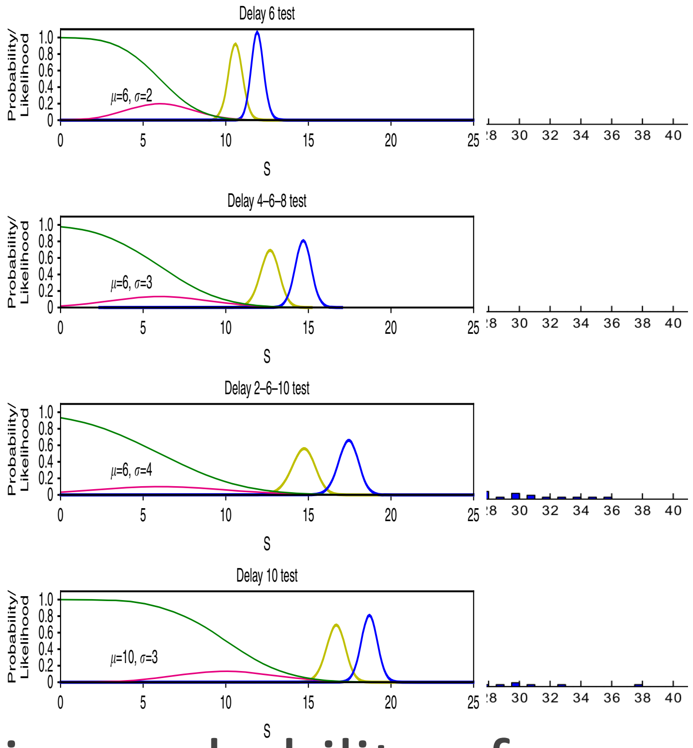
nature COMMUNICATIONS (2018)

Reward probability and timing uncertainty alter the effect of dorsal raphe serotonin neurons on patience

Katsuhiko Miyazaki[1], Kayoko W. Miyazaki[1], Akihiro Yamanaka[2], Tomoki Tokuda[3], Kenji F. Tanaka[4] & Kenji Doya[1]

- Serotonin stimulation facilitates waiting when...
  - reward delivery is certain
  - reward timing is uncertain



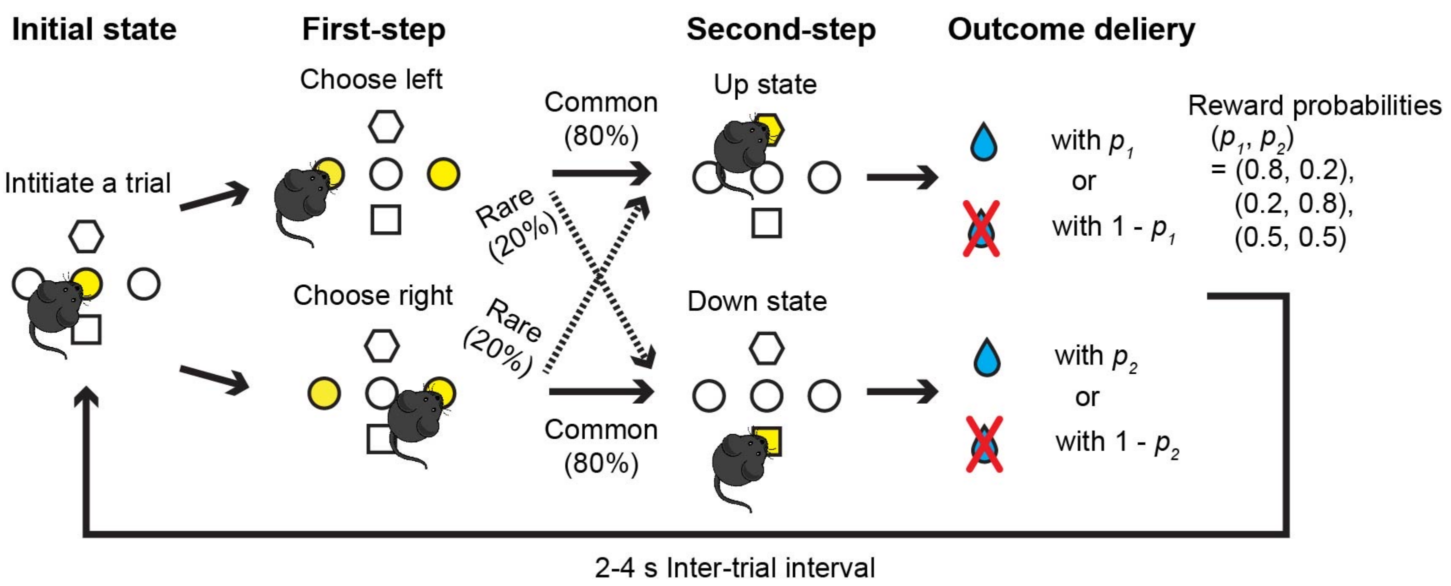


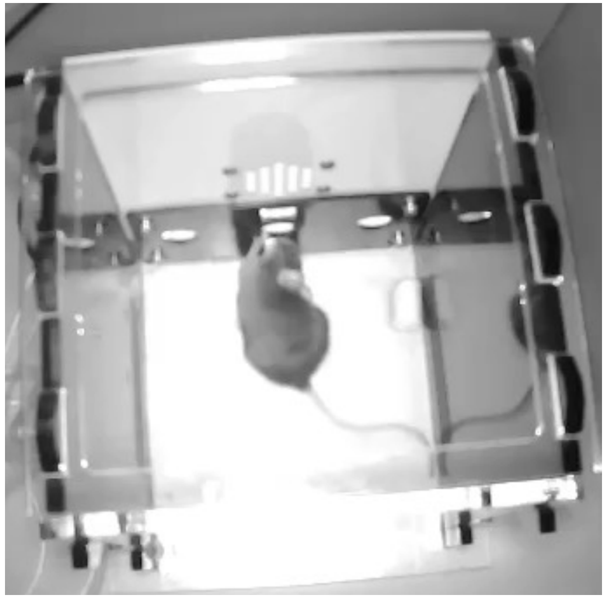
- Reproduced by assuming 5-HT enhances prior probability of reward.

Serotonin for Model-based RL?

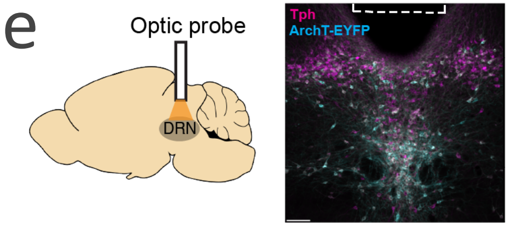
Masakazu Taira

Two-step task for mice (Akam et al. 2020)



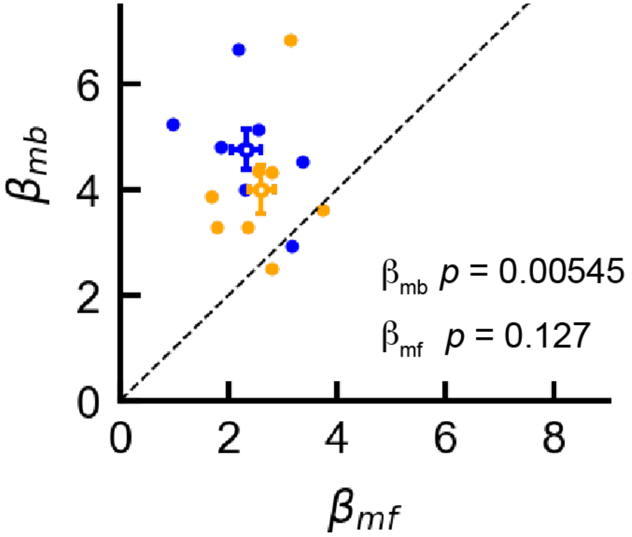


- Tph2-ArchT mice



- Hybrid model

$$Q_{net}(a) = \beta_{\mathrm{mf}} Q_{mf}(a) + \beta_{mb} Q_{mb}(a)$$

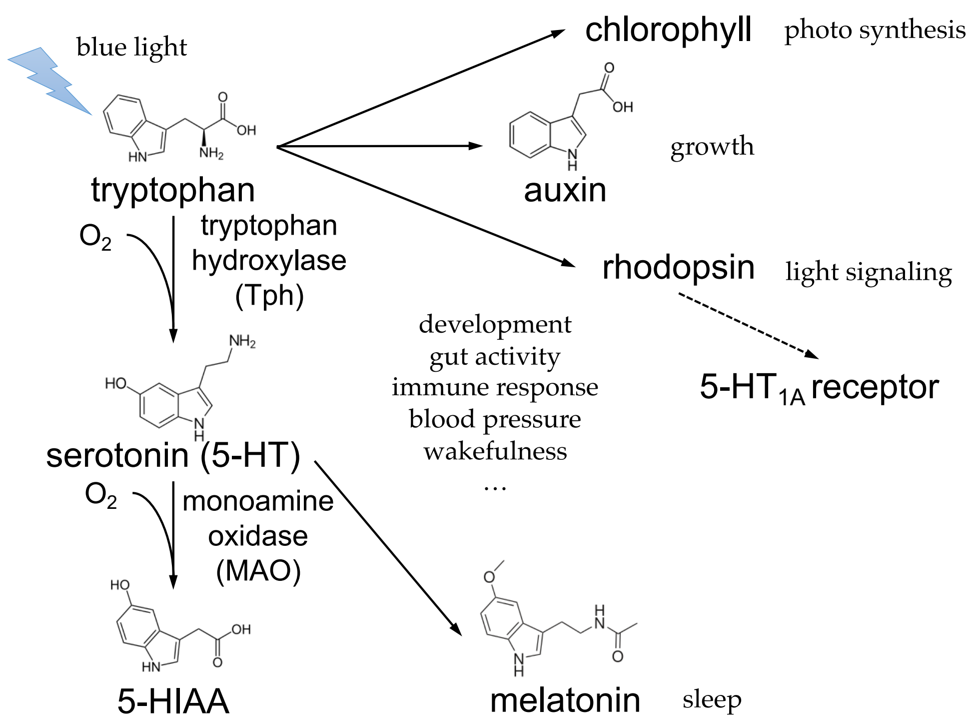


Serotonin Signals Available Time and Resources?

Serotonergic modulation of cognitive computations

Kenji Doya, Kayoko W Miyazaki and Katsuhiko Miyazaki (2021)

Current Opinion in Behavioral Sciences



	Less time	More time
Development	stay	grow
Energy metabolism	utilize	save
Action vigor	spurt	relax
Risk taking	gamble	safe
Threat response	freeze, panic	cope, avoid
Social decision	selfish	cooperative
Learning rate α	fast	slow
Exploration β	exploit	explore
Temporal discounting γ	steep	slow
Eligibility trace λ	short	long
TD error component δ	immediate	predictive
Decision strategy	model-free	model-based
Search	narrow, shallow	wide, deep
Sensory perception	biased to prior	more evidence
Confidence in reward	low	high

Multidisciplinary Frontier Brain and Neuroscience Discoveries

Brain/MINDS 2.0

The Brain/MINDS 2.0 program was launched on March 5, 2024.

Until the official website opens, get updated information about the program here!

Go to the Japanese page.

Topics

2024/2/21：The Brain/MINDS 2.0 is a large-scale national research program in the field of brain science in Japan. The Japan Agency for Medical Research and Development (AMED) selected the "Core Organization" of Brain/MINDS 2.0.

Principal Research Institution: RIKEN

Subsidiary Research Institution: The University of Tokyo, Kyoto University, QST, NCNP, NIPS, ATR, and OIST

2024/5：Overview of the Brain/MINDS 2.0 Core Organization has opened.

Brain/MINDS 2.0: Digital Brain Development

What is a Digital Brain?

Integration of anatomical/physiological/behavioral data into a mathematical model to reproduce brain dynamics and functions

Reproduce brain functions in perception, motion, cognition,...

➢ Contribution to neuroscience and brain-inspired AI

Predict the effects of changes in brain areas, cells, molecules,...

➢ Contribution to pathology and diagnosis/therapy/prevention.

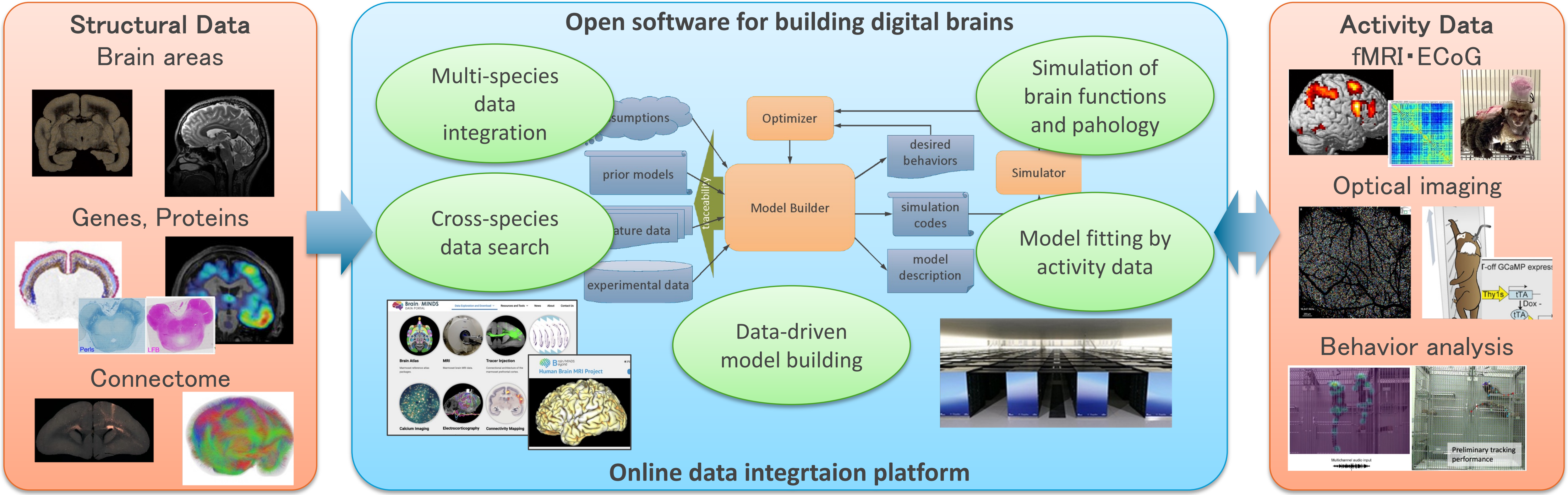
Goals

Open software for building digital brains

Online platform for model building and simulation

Targets of Applications

- Networks for reinforcement learning/Bayesian inference
- Prediction of pathogenic protein propagation
- Therapy planning by psychiatric disorder model

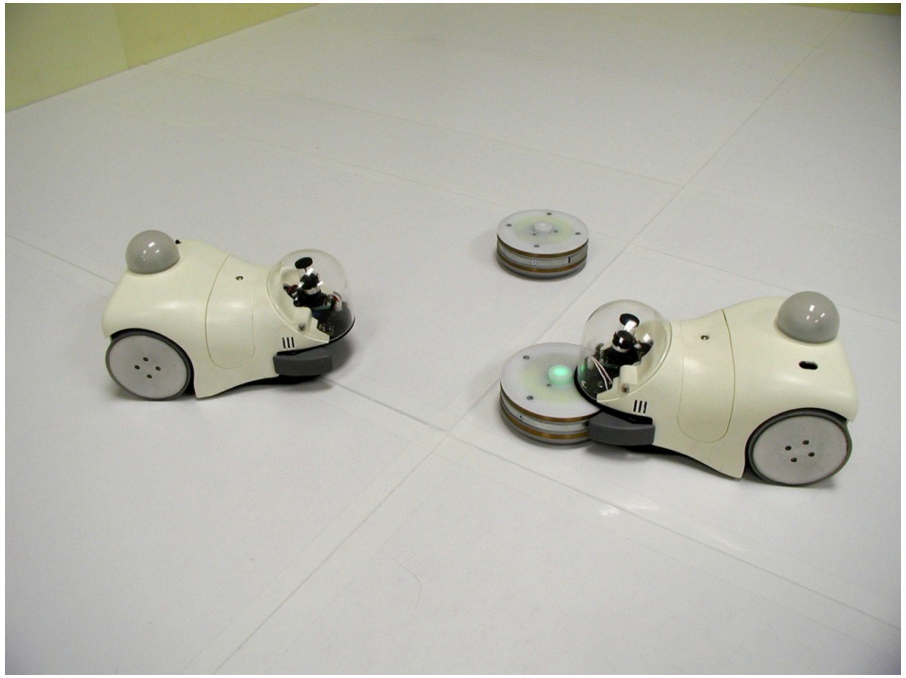


Cyber Rodent Project (Doya & Uchibe, 2005)

What is the origin of rewards?

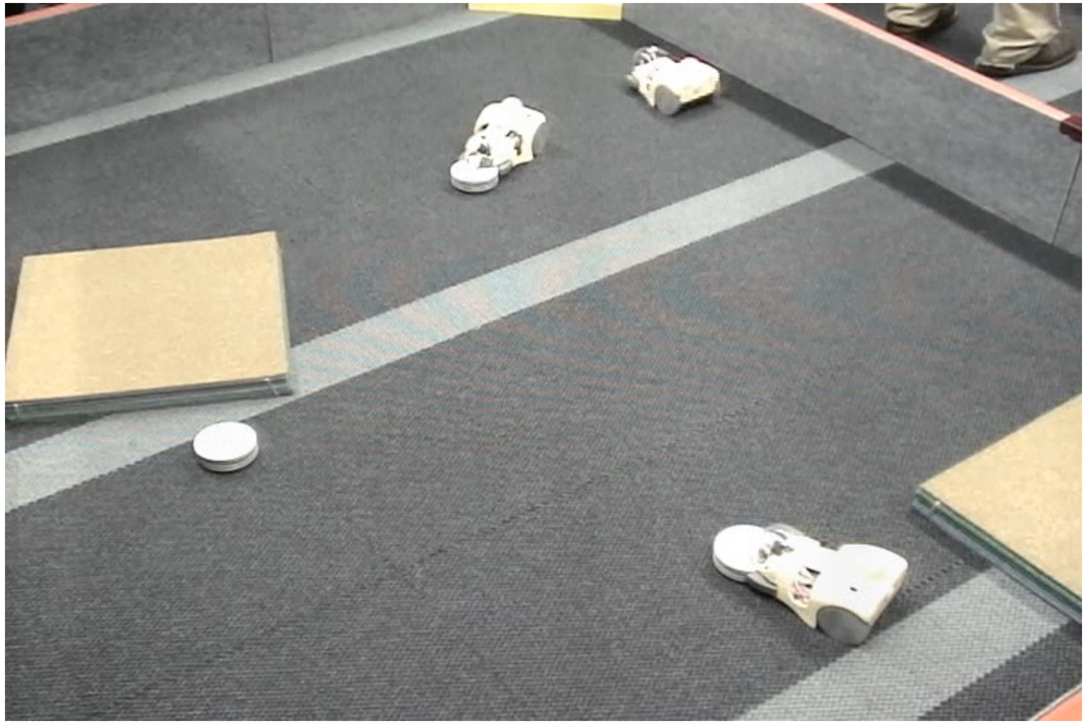
Robots with same constraint as biological agents

- Self-preservation
  - capture batteries
- Self-reproduction
  - exchange programs through IR ports

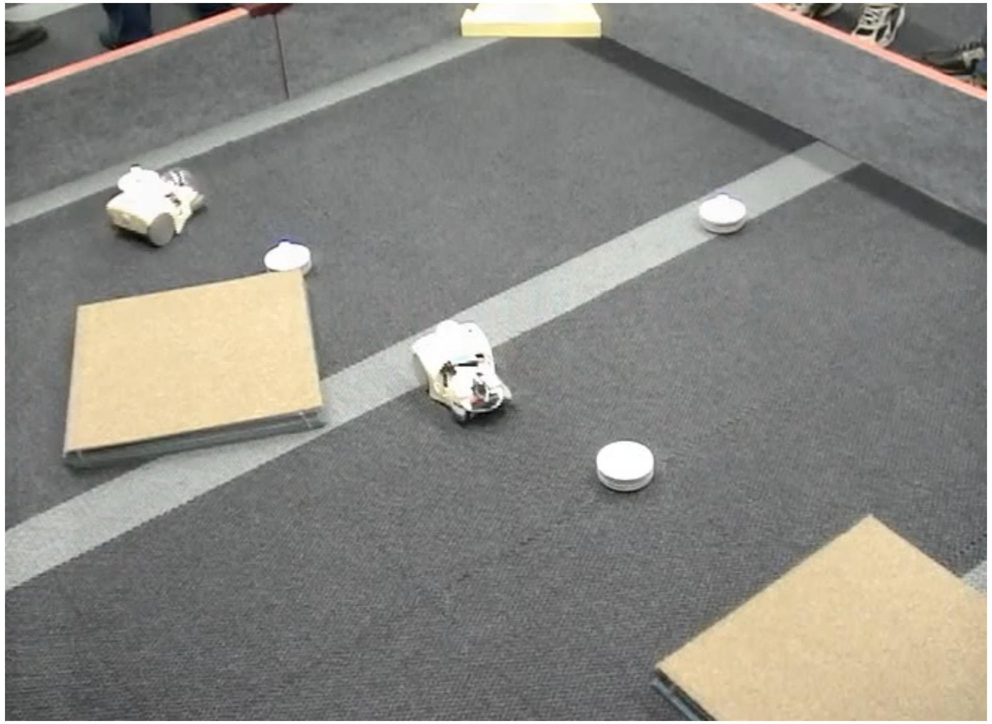


Learning to Survive and Reproduce

- Catch battery packs
  - survival



- Copy ‘genes’ by IR ports
  - reproduction, evolution

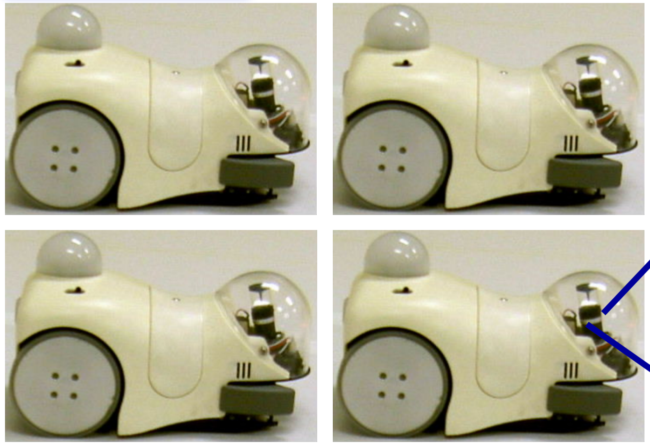


(Doya & Uchibe, 2005)

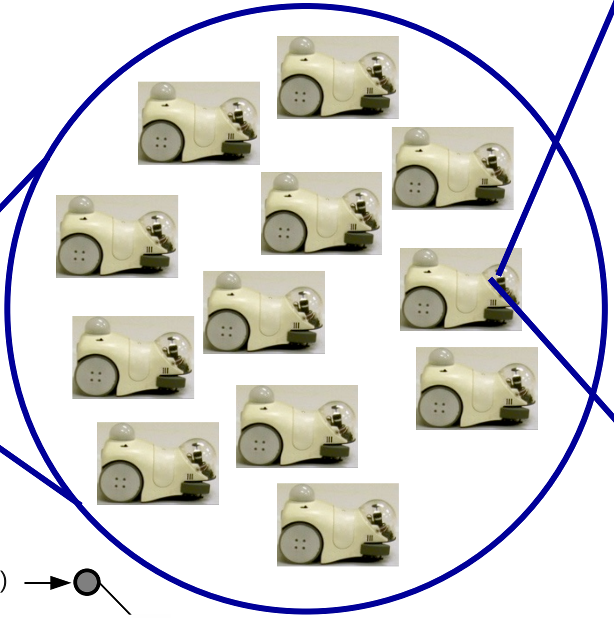
Embodied Evolution (Elfwing et al., 2011)

Population

Robots



Virtual agents
15-25



Genes

Weights for top layer NN

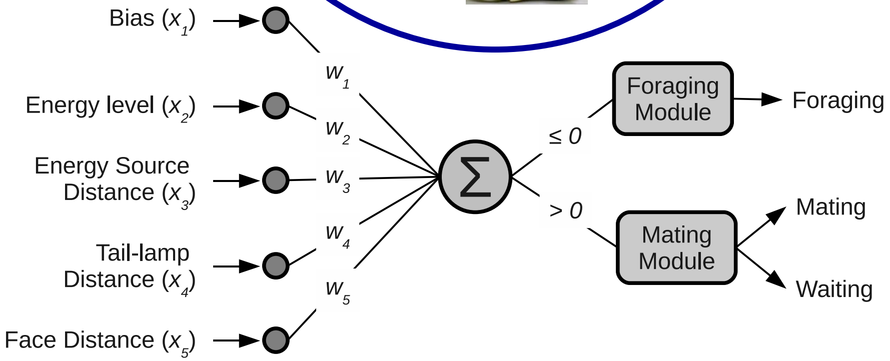
$w_1, w_2, \ldots, w_n$

Weights shaping rewards

$v_1, v_2, \ldots, v_n$

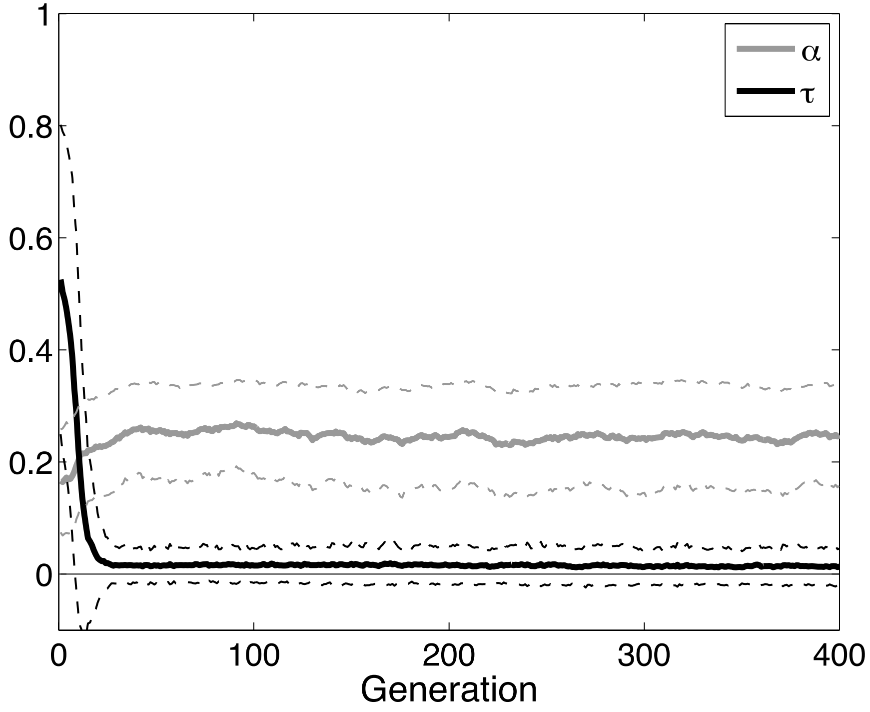
Meta-parameters

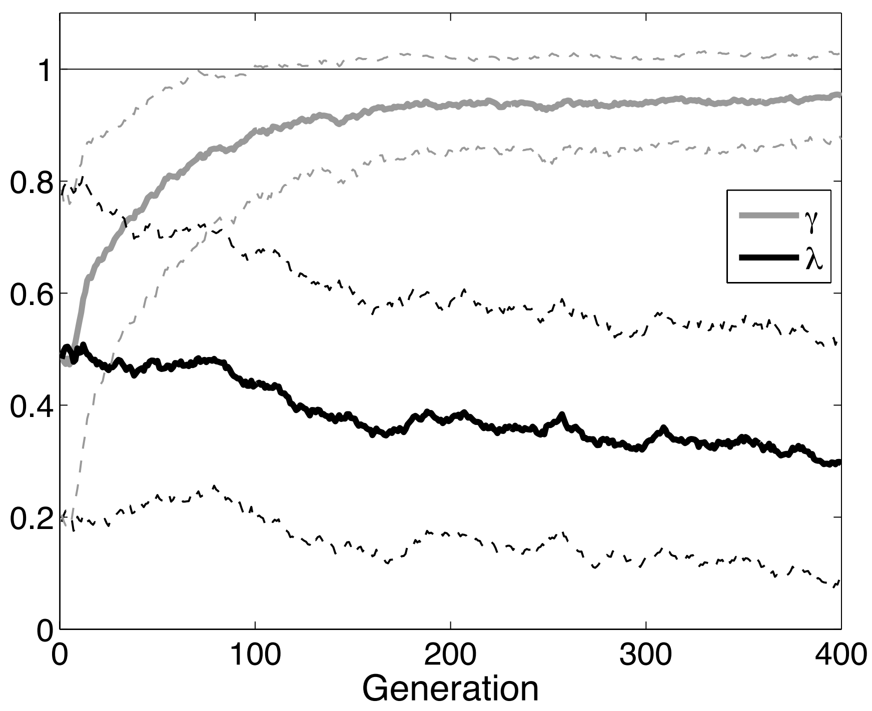
$\alpha \gamma \lambda \tau_k \tau_0$



Evolution of Meta-Parameters

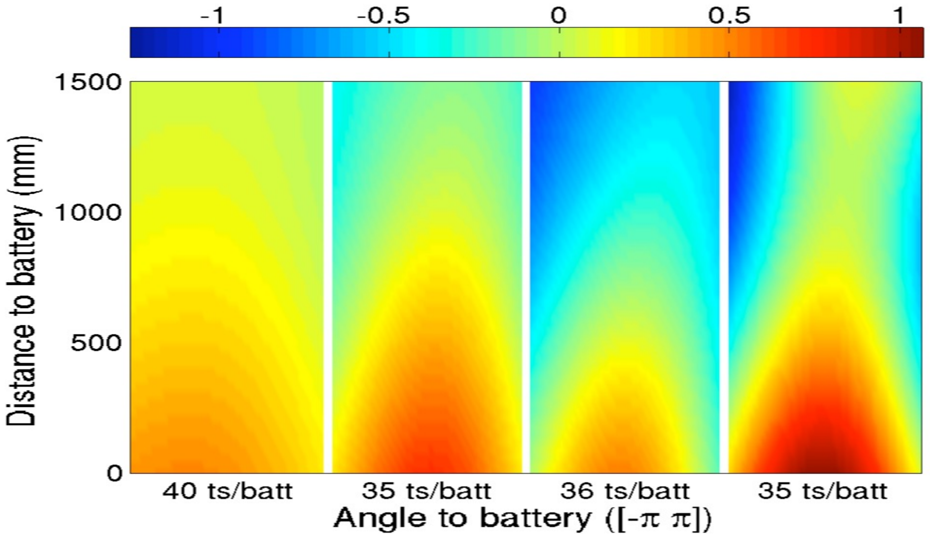
- Learning rate α
- Exploration temperature τ
- Temporal discount factor γ
- Eligibility trace decay factor λ



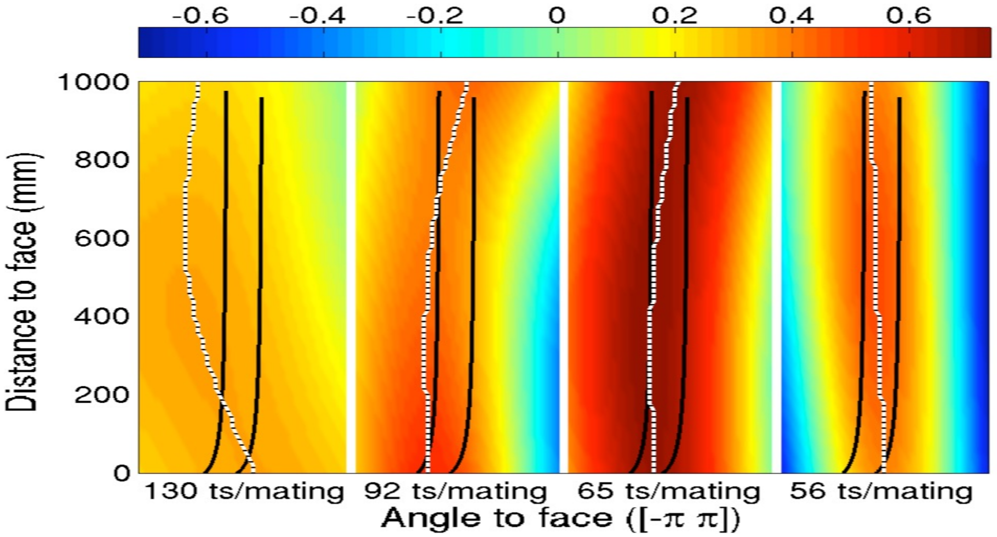


Evolution of Shaping Rewards

- Vision of battery



- Vision of face

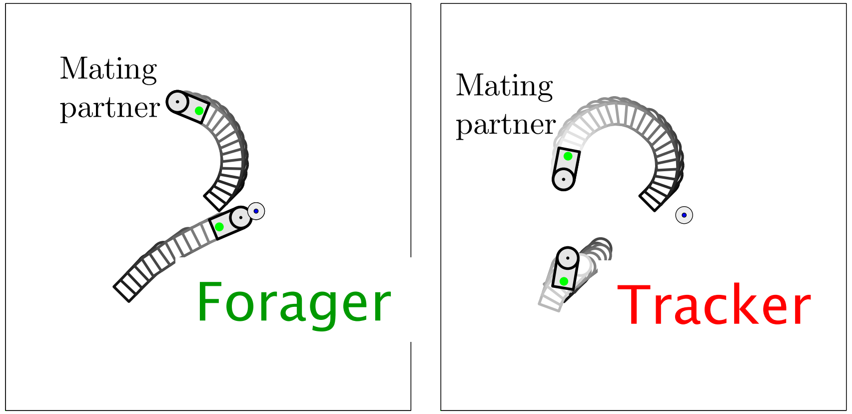


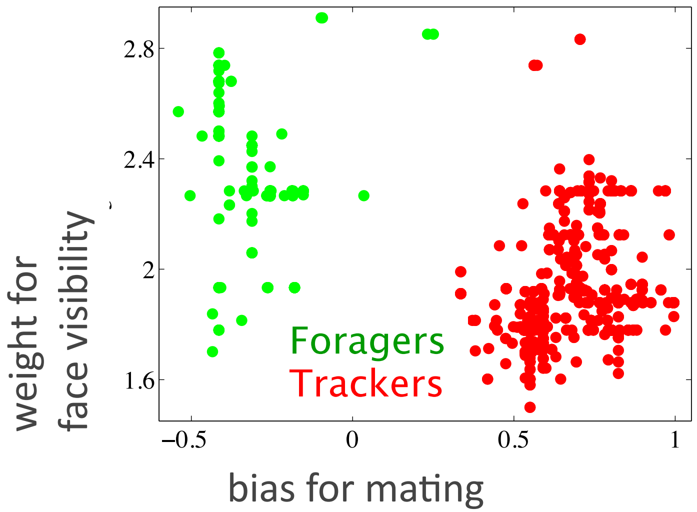
(Elfwing et al., 2011)

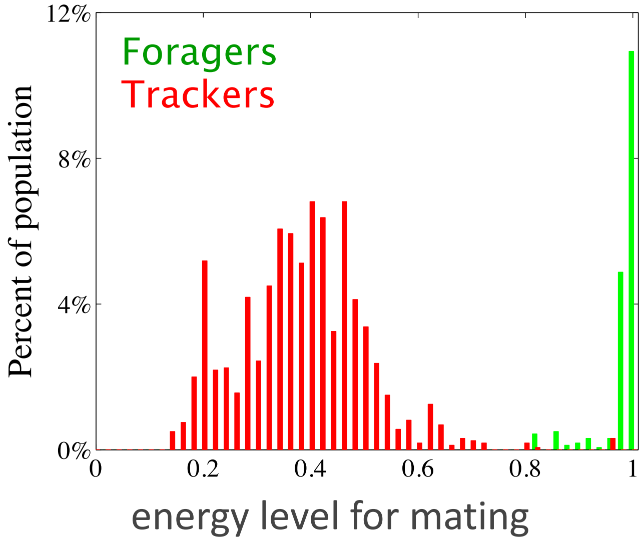
Polymorphism within Colony

(Elfwing et al. 2014)

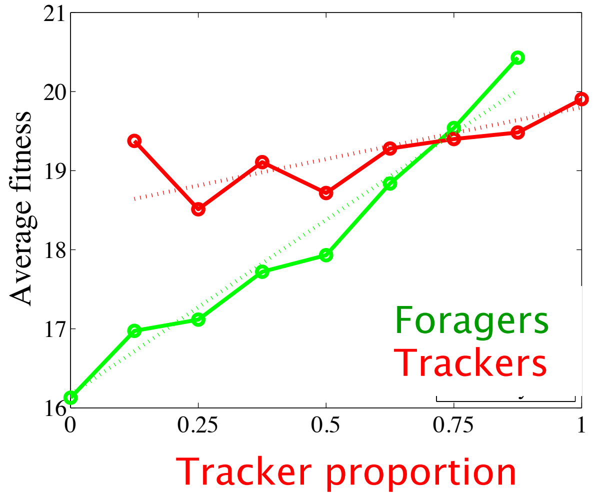
- Foragers and Trackers







- Evolutional stability



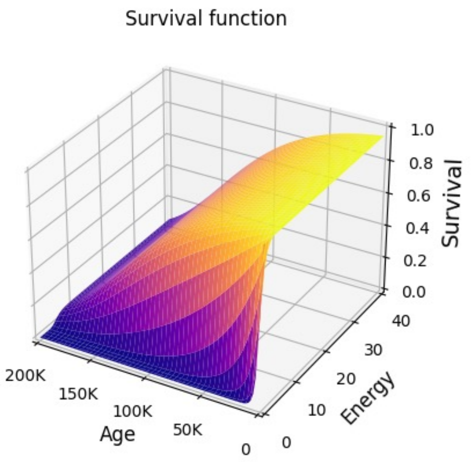
Evolution of Primary Rewards

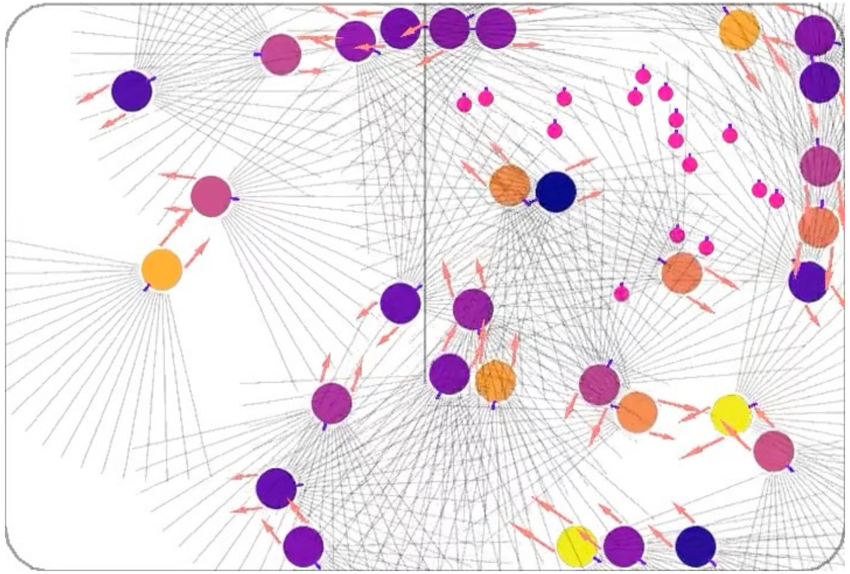
(Yuji Kanagawa, ALIFE 2024)

Reproduction Model

- age t
- energy e

- Death rate $h(t,e)$
- Birth rate $b(e)$

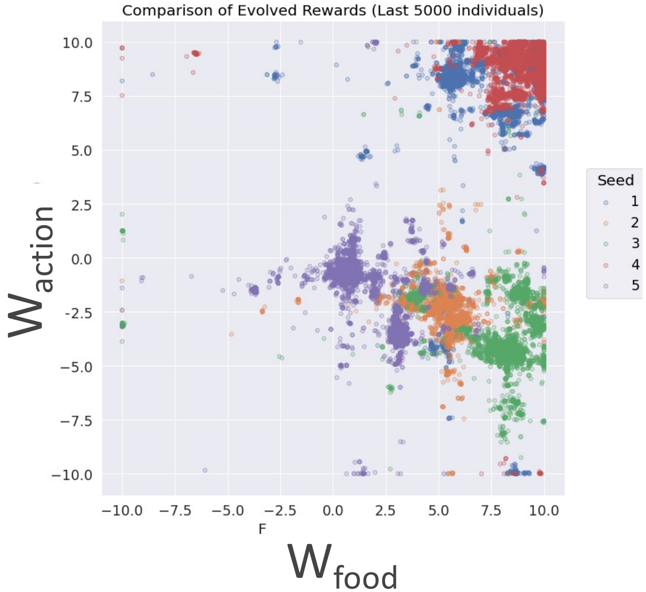




Learning by Reward Function

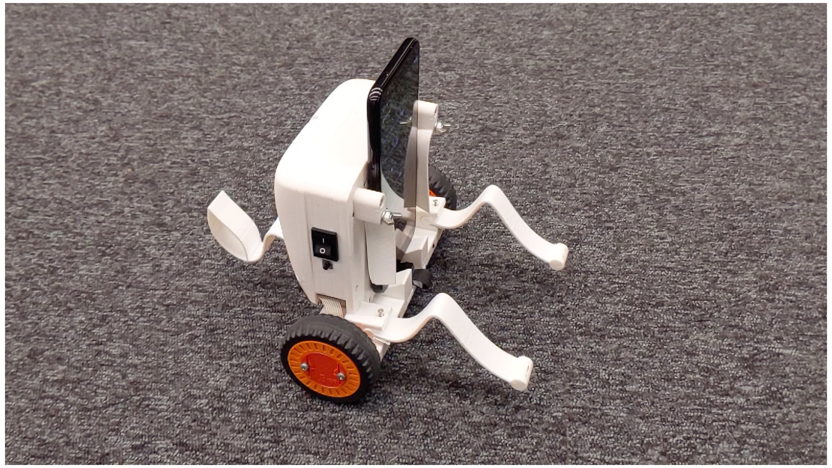
$$r = r_{\text{agent}} + r_{\text{food}} + r_{\text{wall}} + r_{\text{action}}$$

Evolution of Reward Function

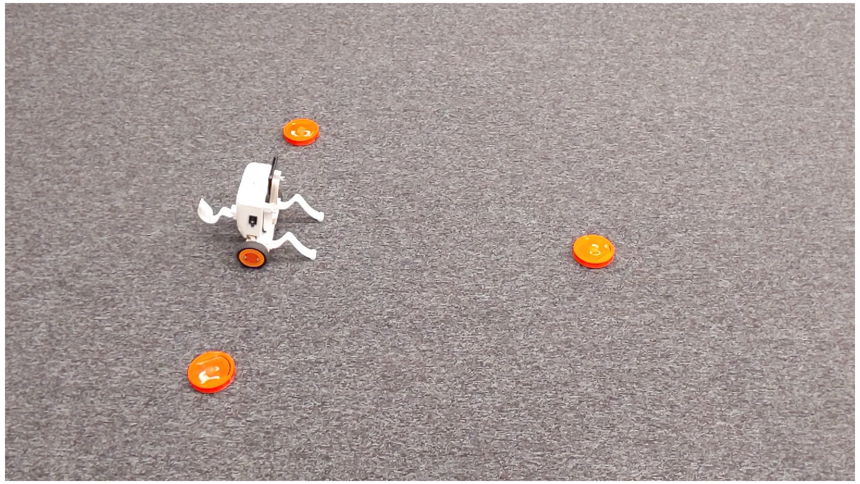


Smartphone Robot Project

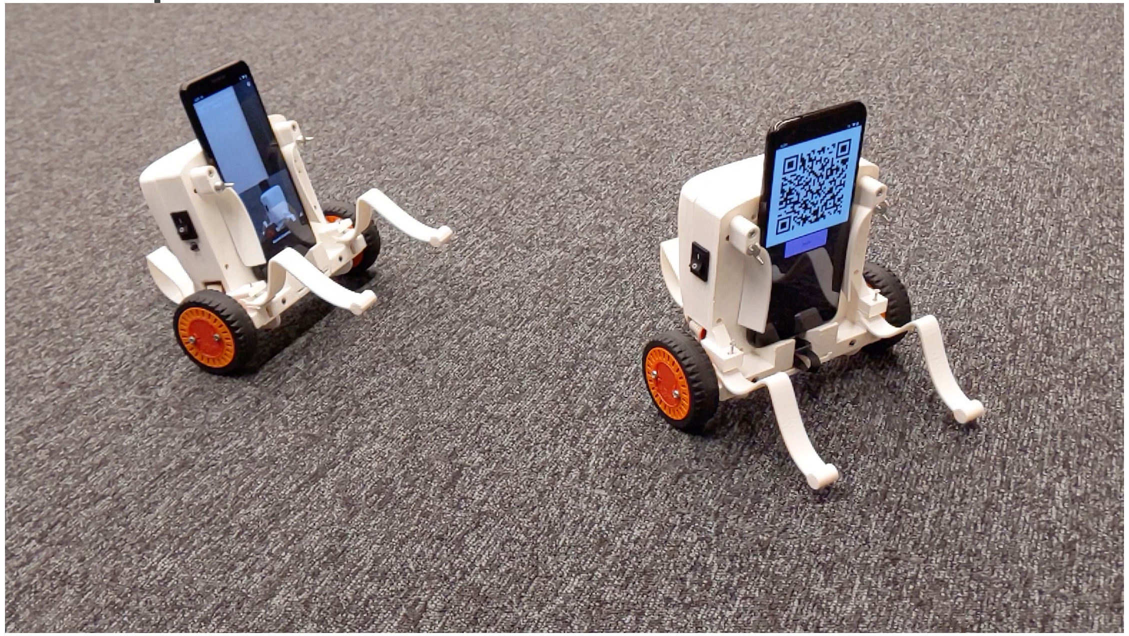
- Motor control



- Survival



- Reproduction



- Learning models of world and others
- Meta-learning
- Evolution of rewards and curiosity
- …

Danger of Autonomous AI?

AI agents can find new goals and try them out

- Creating novel science, technology, culture, industry..

Assessment and control of dangers

- Overruns, side effects
- *Exploitation by individuals/groups with ambition/hatred*

Learn from human societies

- Humans are the most dangerous species on earth
- Democracy: don’t give unlimited power to a person/group
  - election, term limit, separation of powers
  - antimonopoly, right to strike, information disclosure

Peer reviewing among open-sourced, explainable AI agents

IEEE WCCI 2024

Home About Authors Program Travel Sponsors

World Congress on Computational Intelligence (WCCI) 2024

AIガバナンス公開フォーラム

Open Forum on AI Governance

image by DALL-E

2024年6月30日（日）9:20 – 18:00　パシフィコ横浜

6月30日から横浜で開催される計算知能国際会議 (WCCI 2024) は人工知能 (AI) に関する今年アジアで最大規模の学会です。AIのもたらす危険性が議論され規制が進む中、AIの開発者、利用者、政策立案者を集めた公開フォーラムを開催します。

招待講演者

ヨシュア・ベンジオ

村上明子

スチュアート・ラッセル

バネッサ・ニューロック他

無料の事前登録により会場またはオンラインでご参加いただけます。発表は英語で日本語のAI翻訳を提供する予定です。幅広く市民、学生の皆さんの参加をお待ちしています。詳細はwebサイト https://groups.oist.jp/ja/ncu/event/wcci-forum をご参照ください

主催：IEEE、国際神経回路学会、日本神経回路学会、他

Acknowledgements

- Striatum recording
  - **Makoto Ito (Progress Technology)**
  - **Tomohiko Yoshizawa (Tamagawa U)**
  - **Charles Gerfen (NIH)**
  - Kazuyuki Samejima (Tamagawa U)
  - Minoru Kimura (Tamagawa U)
- Human fMRI/behavior
  - **Alan Fermin (Tamagawa U)**
  - **Takehiko Yoshida (NAIST)**
  - **Saori Tanaka (ATR)**
  - **Nicolas Schweighofer (USC)**
  - **Jun Yoshimoto (NAIST)**
  - Yu Shimizu
  - Tomoki Tokuda (ATR)
  - Shoko Ota
- Serotonin recording/manipulation/modeling
  - **Kayoko W Miyazaki**
  - **Katsuhiko Miyazaki**
  - **Gaston Sivori**
  - **Masakazu Taira (U Sydney)**
  - **Thomas Akam (Oxford U)**
  - **Mark Walton (Oxford U)**
  - **Kenji Tanaka (Keio U)**
  - **Akihiro Yamanaka (Nagoya U)**
- Cortical imaging
  - **Akihiro Funamizu (U Tokyo)**
  - **Bernd Kuhn**
  - **Yuzhe Li**
  - **Sergey Zobnin**
  - **Naohiro Yamauchi**
- Marmoset data analysis
  - Carlos Gutierrez (Softbank)
  - Hiromichi Tsukada (Chubu U)
  - Junichi Hata, Henrik Skibbe, Alex Woodward (RIKEN)
  - Ken Nakae, (NINS)
- Basal ganglia model
  - **Benoit Girard, Daphne Heraiz (Sorbonne)**
  - **Jean Lienard**
- Robotics
  - Jun Morimoto (ATR)
  - **Eiji Uchibe (ATR)**
  - **Stefan Elfwing (ATR)**
  - **Jiexin Wang (ATR)**
  - **Paavo Parmas (Kyoto U)**
  - Kristine Roque
  - **Yuji Kanagawa**
  - Tojoarisoa Rakotoaritina
  - **Christopher Buckley**

Scientific Research in Transformative Areas
Scientific Research on Innovative Areas

Strategic Research Program for Brain Sciences
Brain/MINDS Project
Fugaku Supercomputing Program