OIST Computational Neuroscience Course 2022, June 18

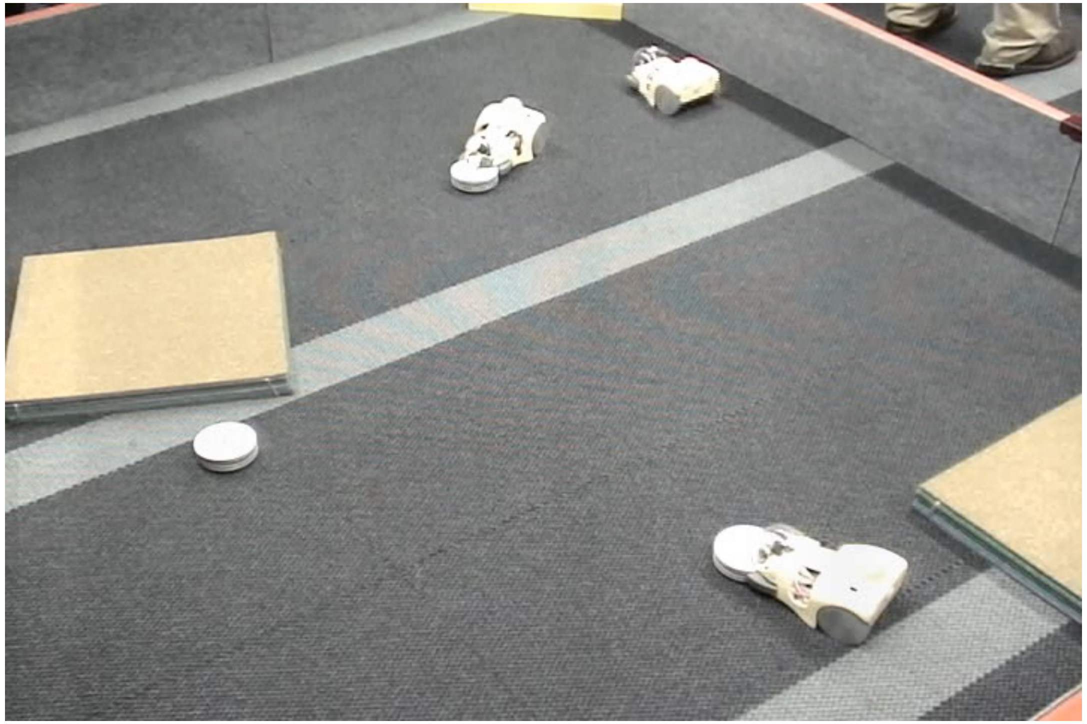
Reinforcement Learning and Bayesian Inference

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

OIST Neural Computation Unit

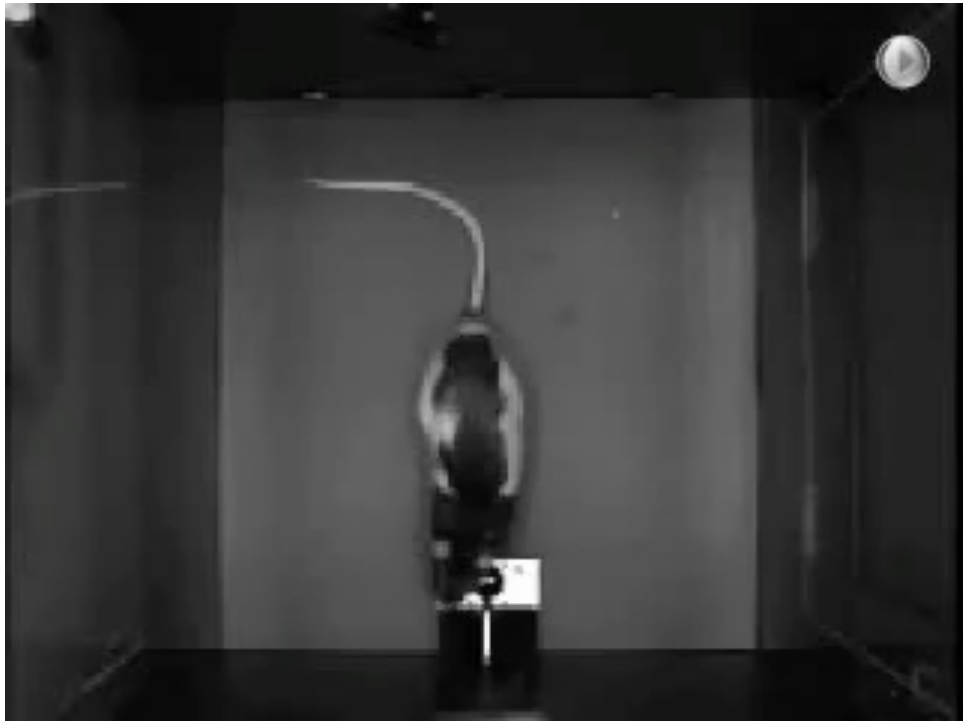
Create flexible learning systems

- robot experiments

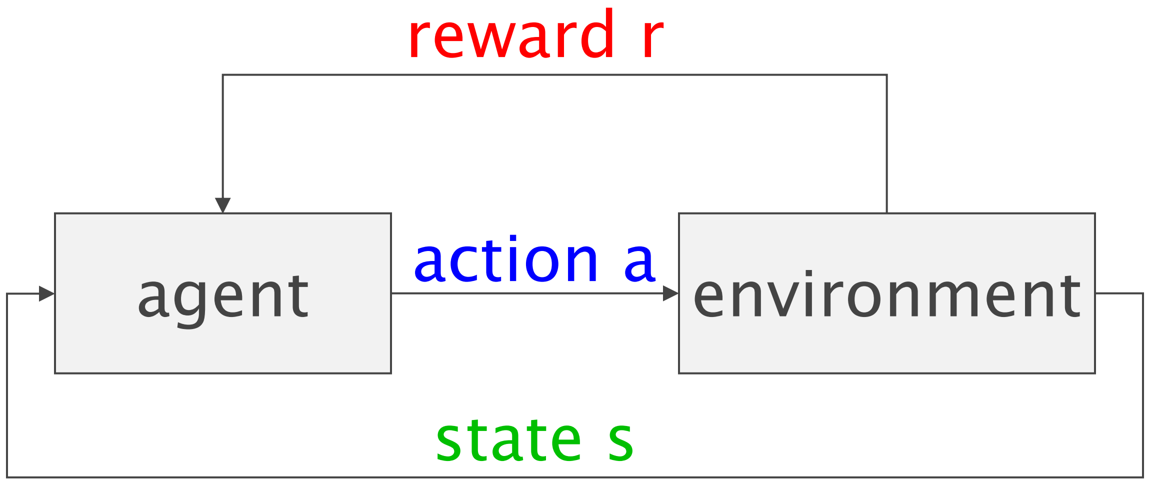


Reveal brain's learning mechanisms

- neurobiology



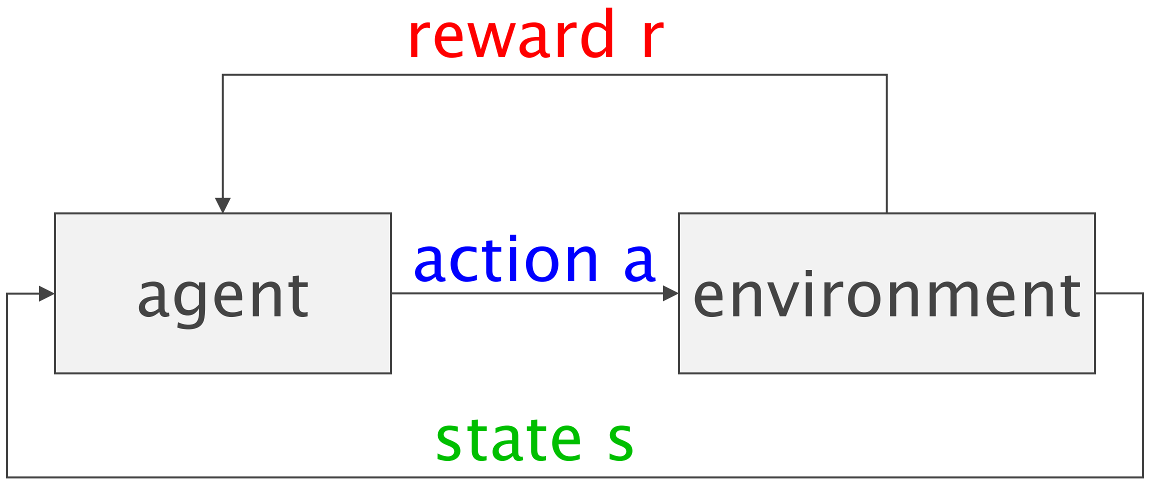
Reinforcement Learning



Learn action policy: s $\rightarrow$ a to maximize rewards

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

Reinforcement Learning



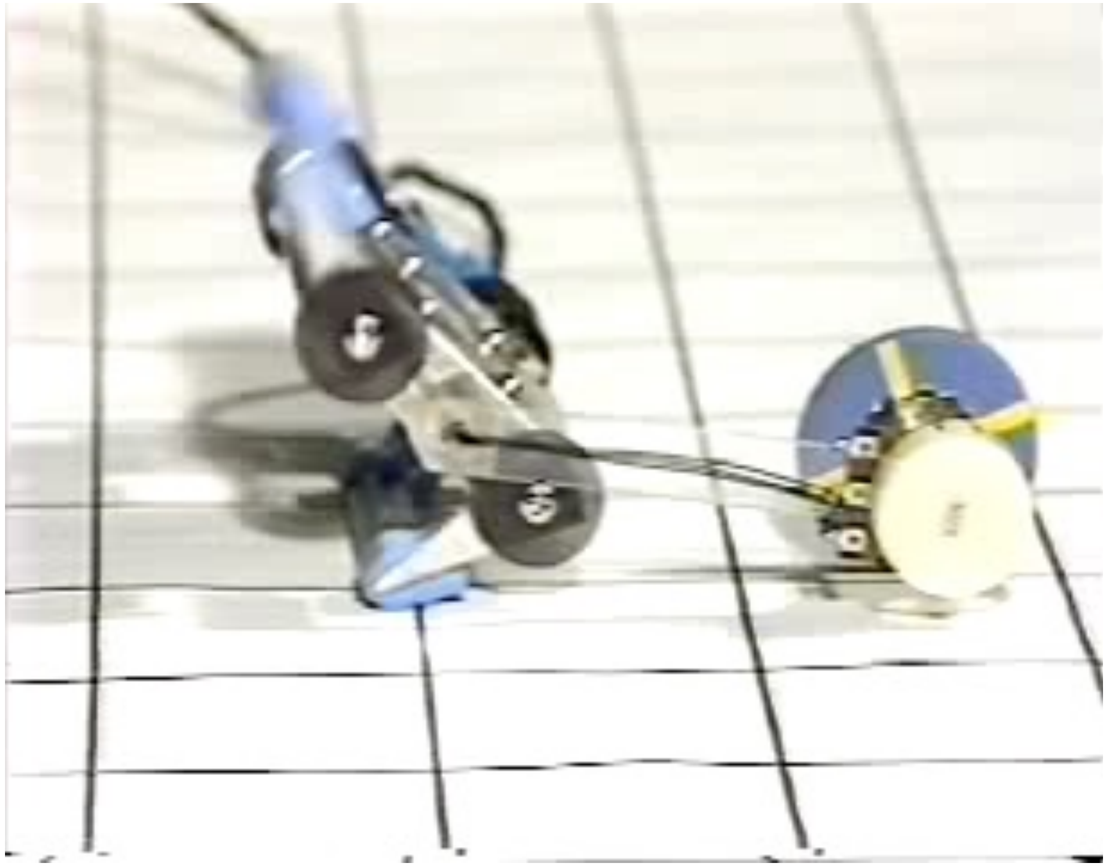
Learn action policy: s → a to maximize rewards

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

Learning to Walk

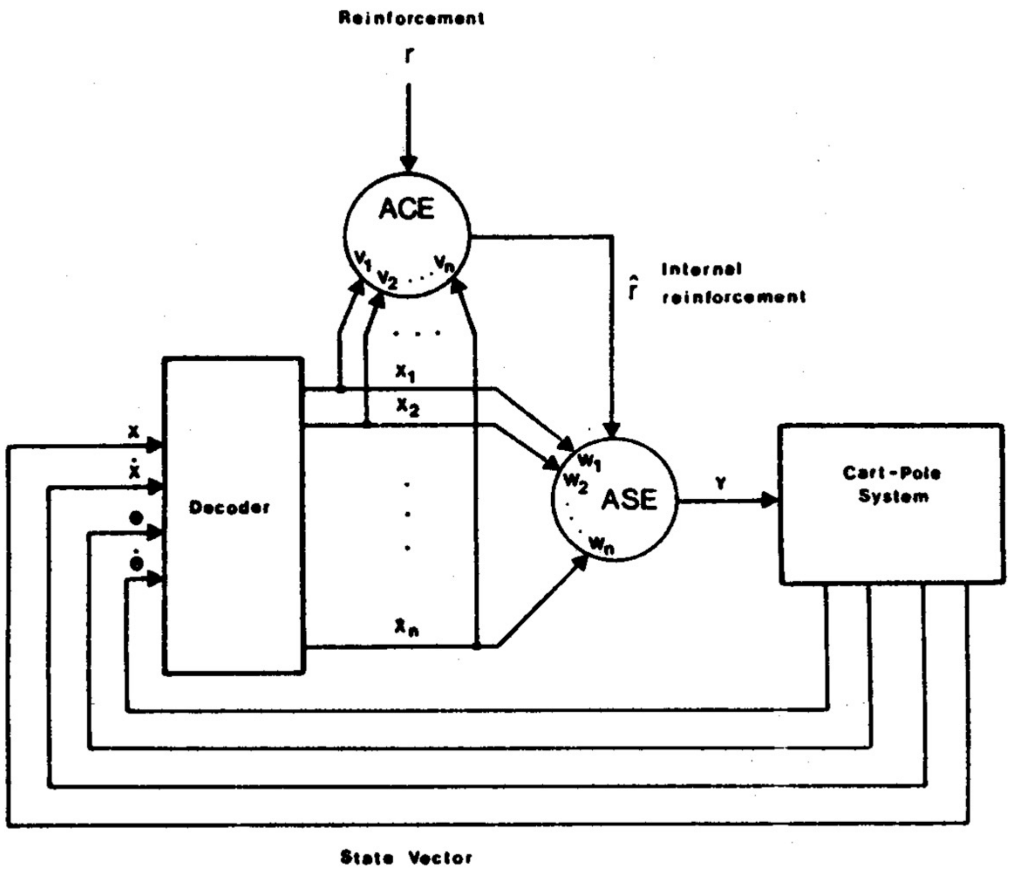
(Doya & Nakano, 1985)

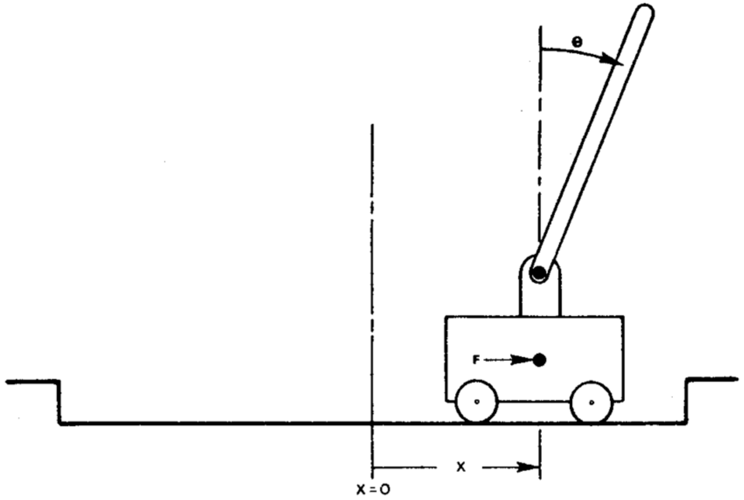
- Explore actions (cycle of 4 postures)
- Learn from performance feedback (speed sensor)

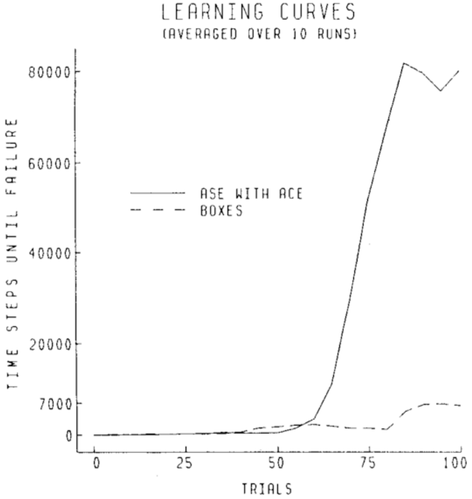


Neuronlike Adaptive Elements That Can Solve Difficult Learning Control Problems

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

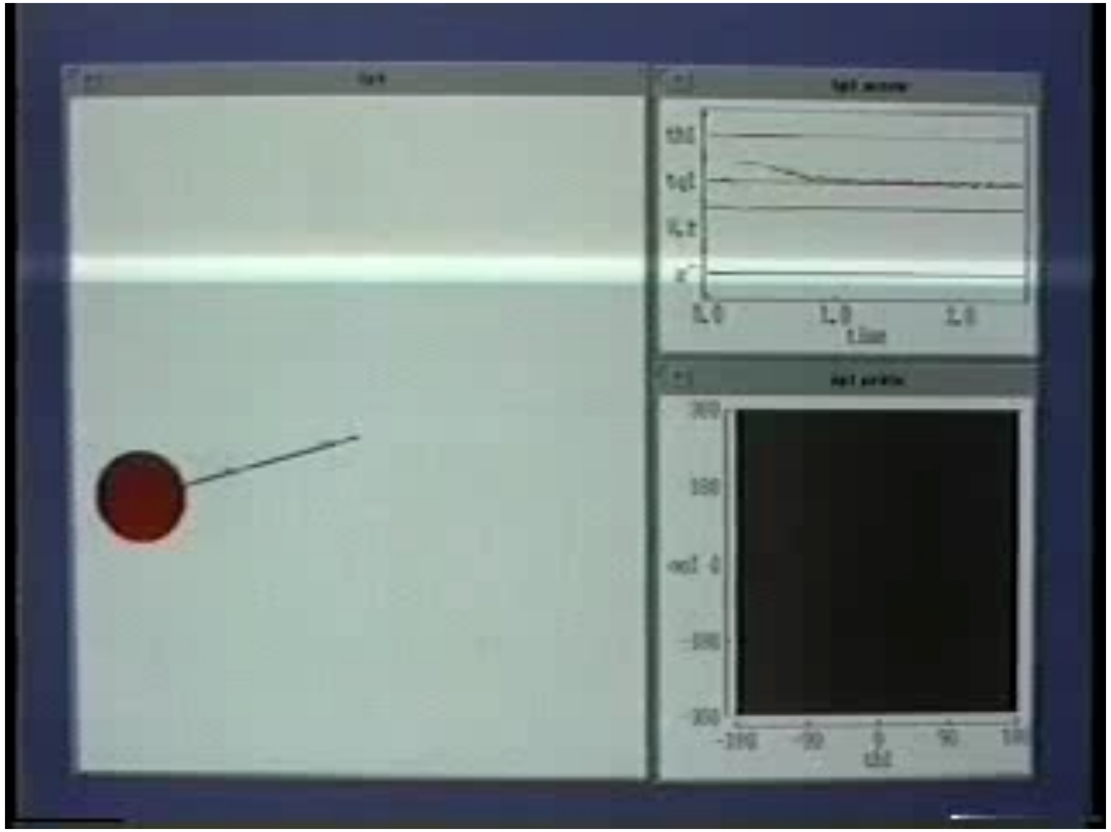






Pendulum Swing-Up

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



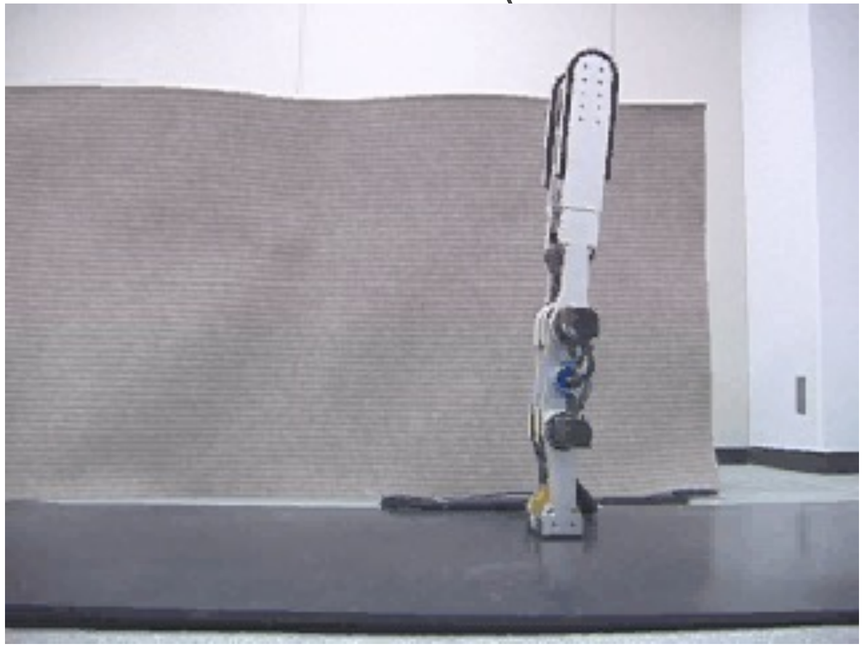
ω

θ

- Value function

Learning to Stand Up

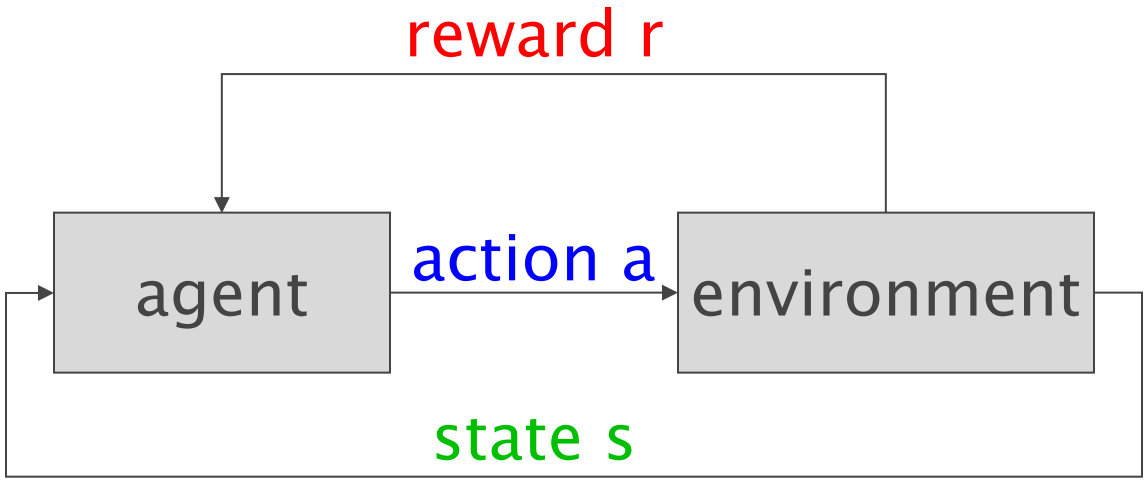
(Morimoto & Doya, 2001)



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

Markov Decision Process (MDP)

- Markov decision process
  - state $s \in S$
  - action $a \in A$
  - 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
  - $V(s(t)) = E[\, r(t) + \gamma r(t+1) + \gamma^2 r(t+2) + \ldots]$

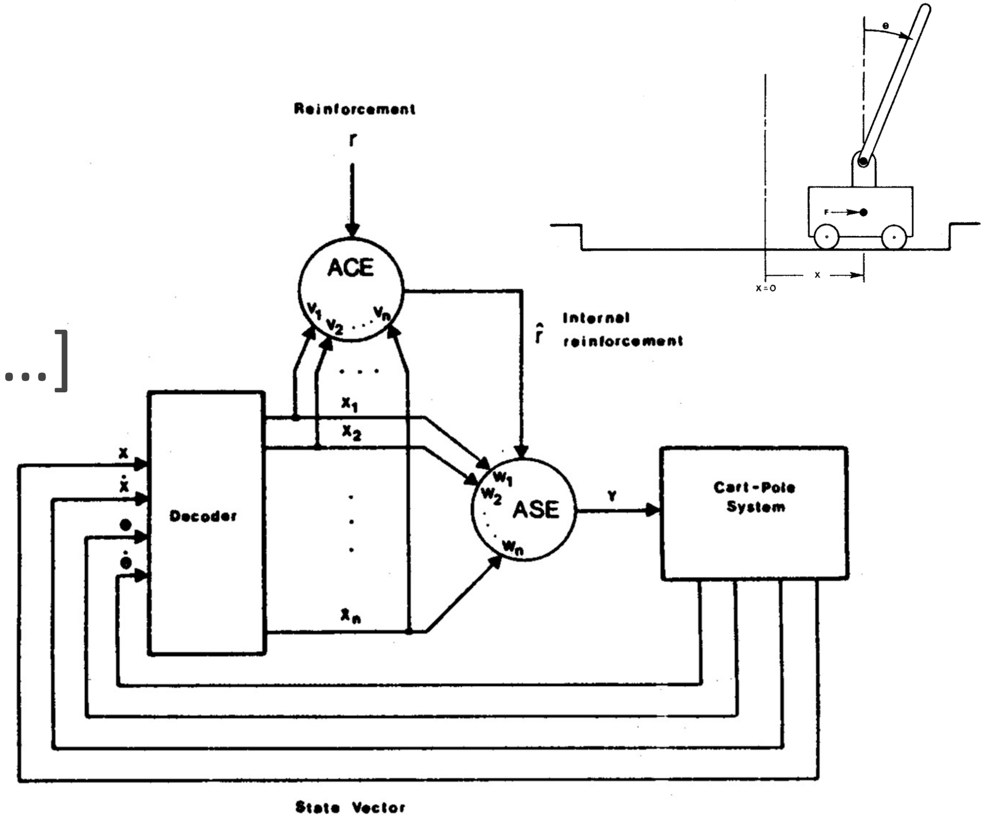
  e.g., $V(s(t);v) = \Sigma_j v_j s_j(t)$

- 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



SARSA and Q Learning

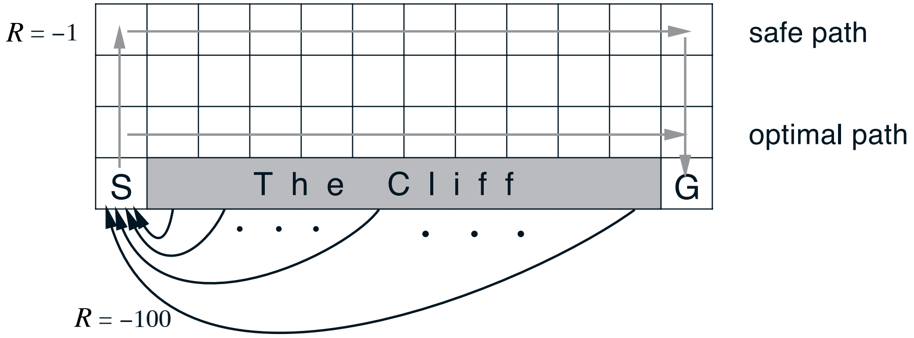
- Action value function
  - $Q(s,a) = E[\, r(t) + \gamma r(t+1) + \gamma^2 r(t+2) \ldots \mid s(t)=s, a(t)=a]$
- Action selection
  - ε-greedy: $a = \mathrm{argmax}_a\, Q(s,a)$ with prob $1-\varepsilon$
  - Boltzman: $P(a_i|s) = \exp[\beta Q(s,a_i)] \,/\, \Sigma_j \exp[\beta Q(s,a_j)]$
- Update by temporal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

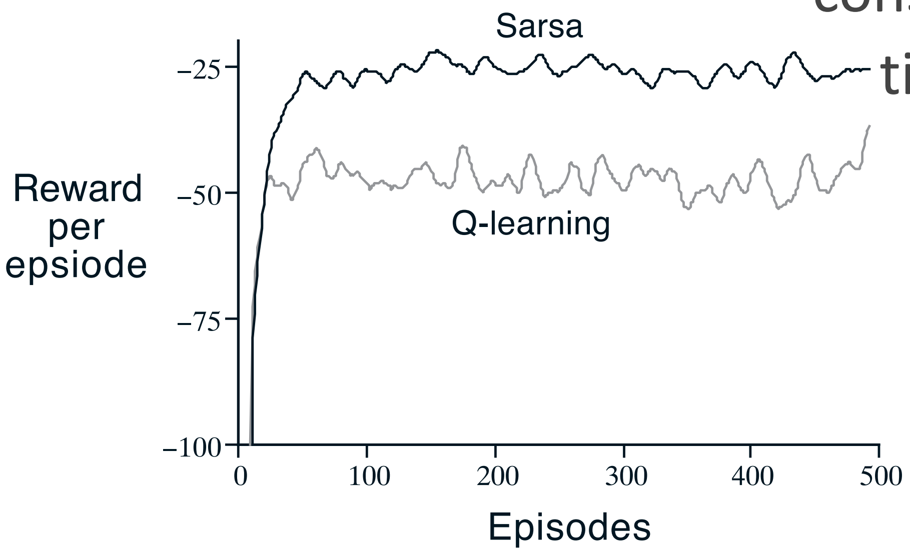
    $\delta(t) = r(t) + \gamma Q(s(t+1),a(t+1)) - Q(s(t),a(t))$
  - Q learning: off-policy

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

SARSA and Q Learning

- Cliff walking task (Sutton & Barto, 1998)



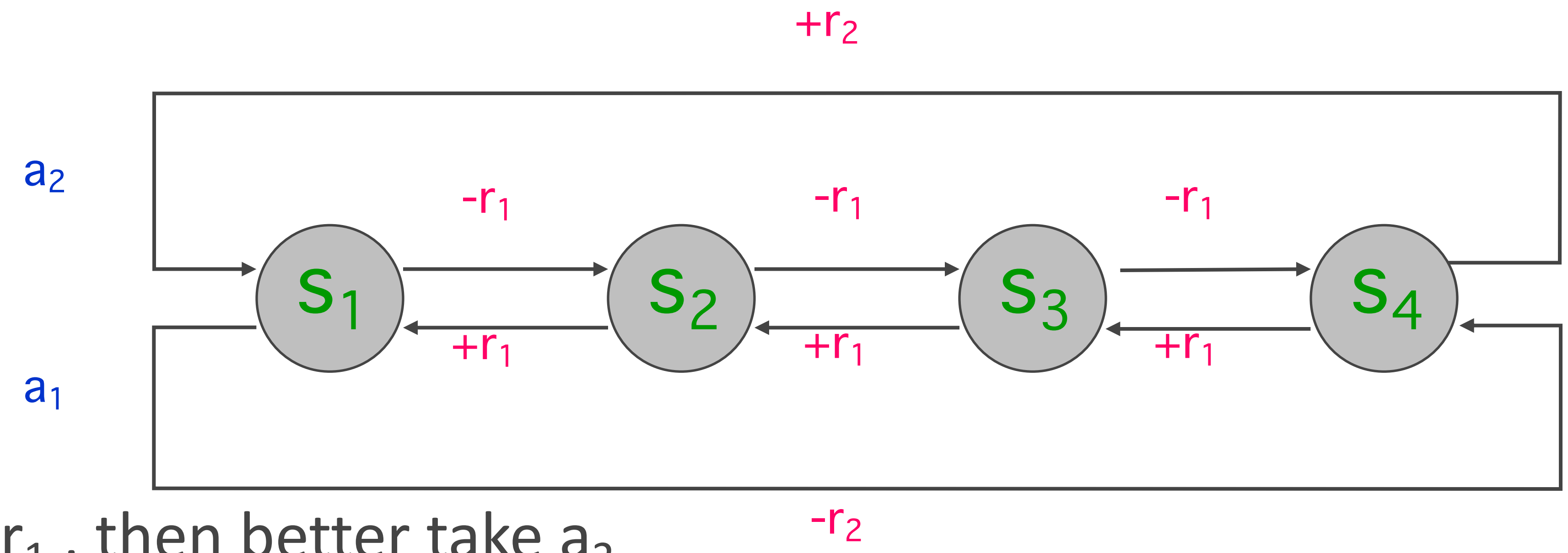


consider exploration

timal greedy policy

"Pain-Gain" Task

- N states, 2 actions

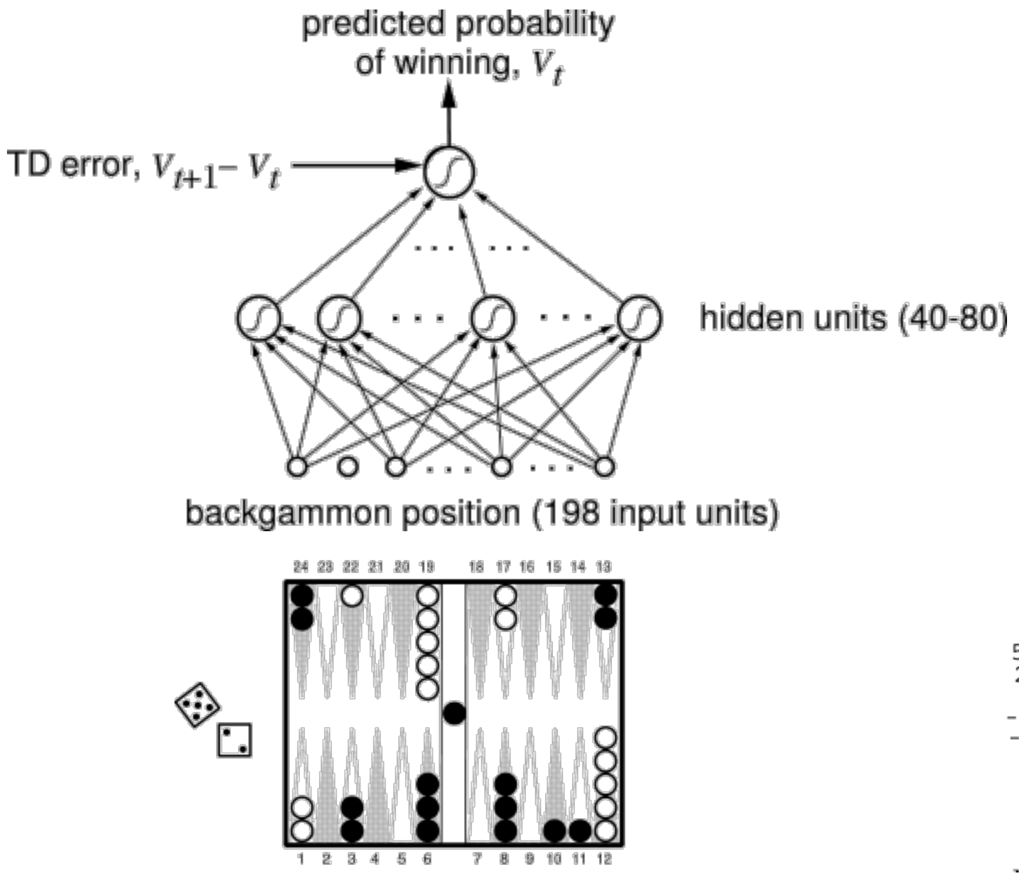


- if $r_2 >> r_1$, then better take a_2

TD Learning and Backprop

- TD Gammon

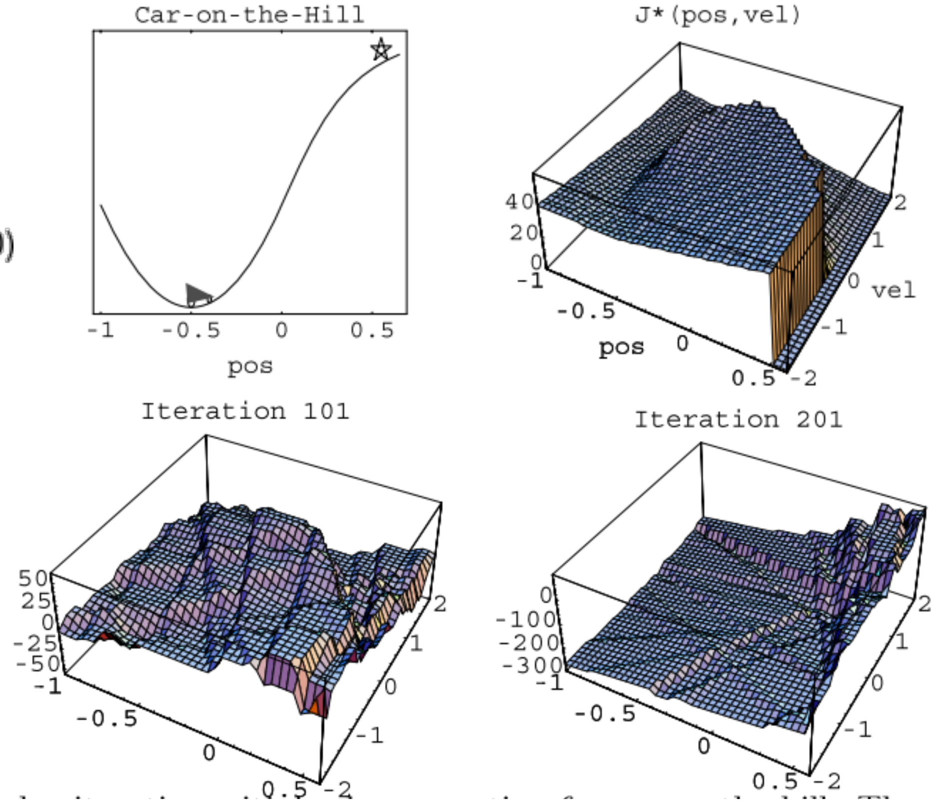
(Tesauro 1992, 1994)



- TD Learning can diverge

(Boyan & Moore, 1995)

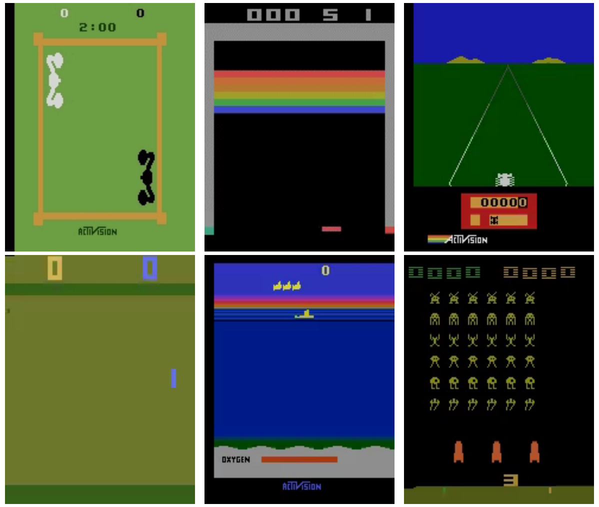
  - $\delta(t) = r(t) + \gamma\, V(s(t+1)) - V(s(t))$

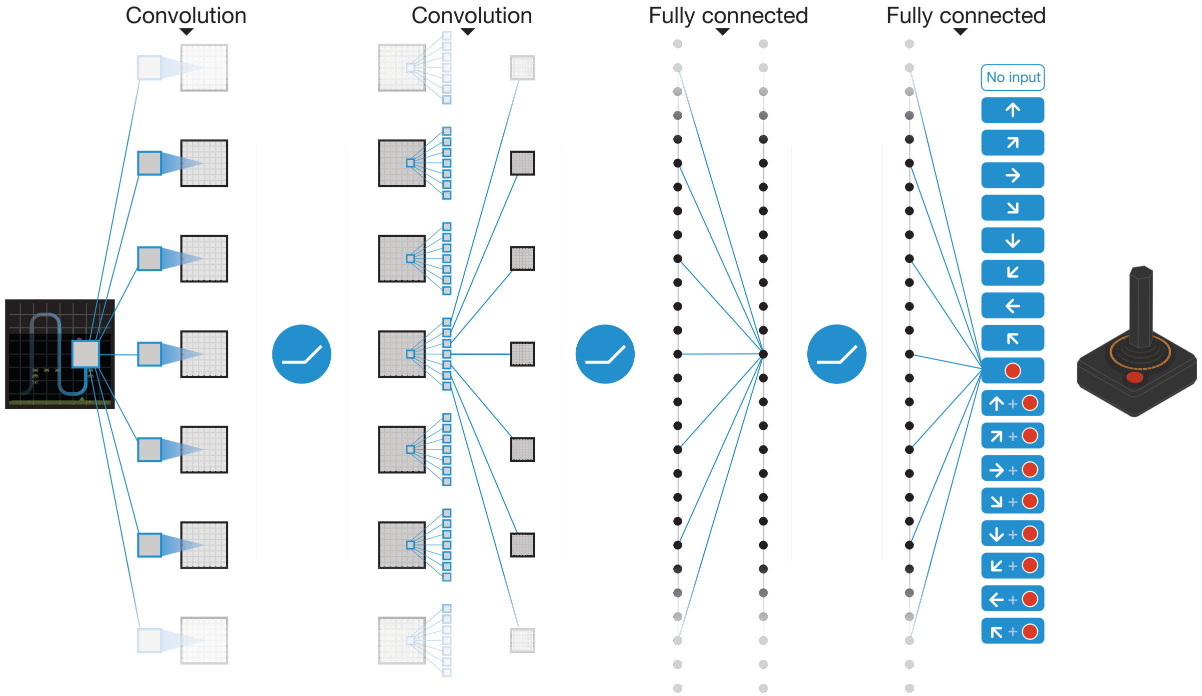


Deep Q-Network

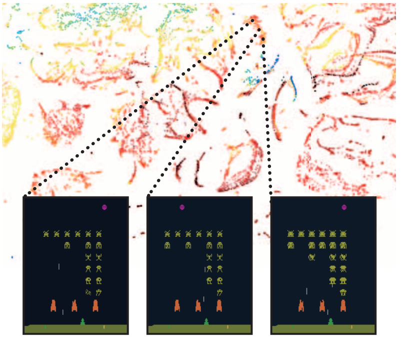
(Mnih et al. 2015)

- Game screen as input





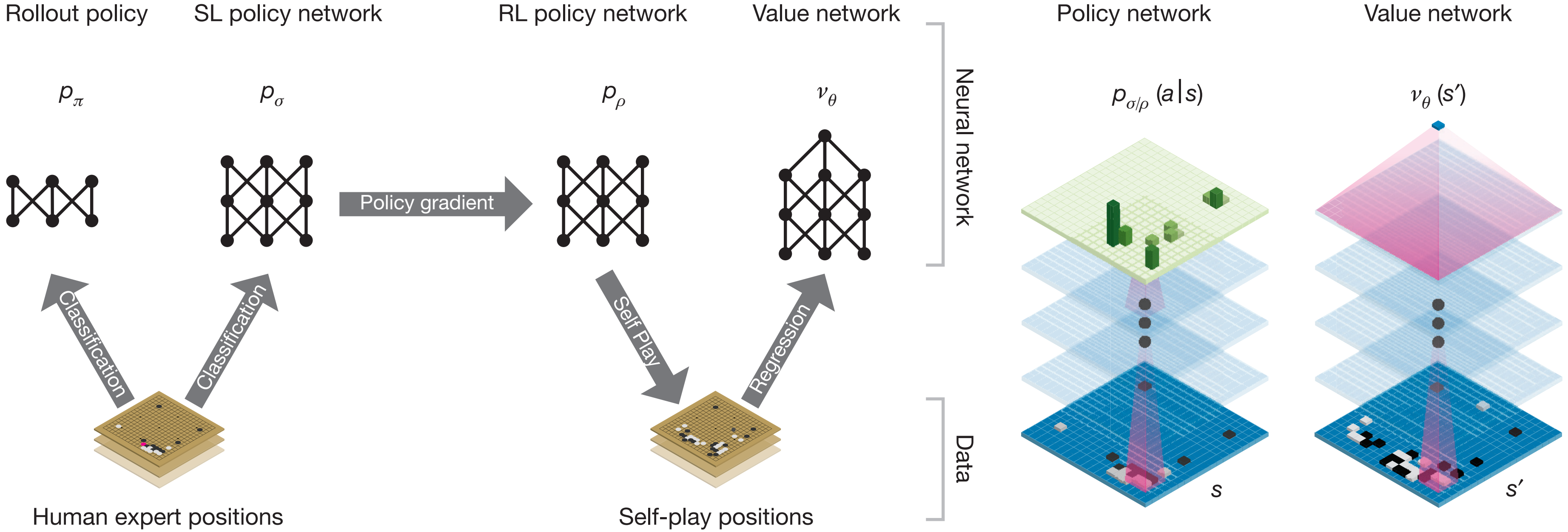
  - *Experience replay*
  - Fixing the *target network*
- DNN captures important features
  - human level in 29/49 Atari games



AlphaGo

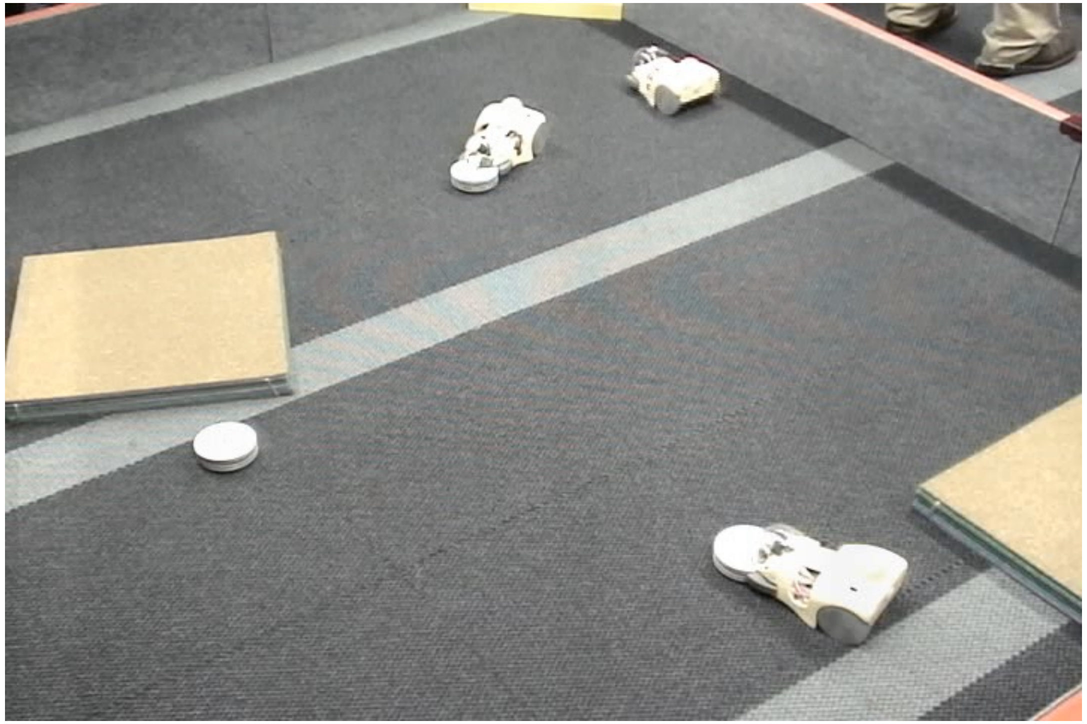
(Silver et al., 2016)

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

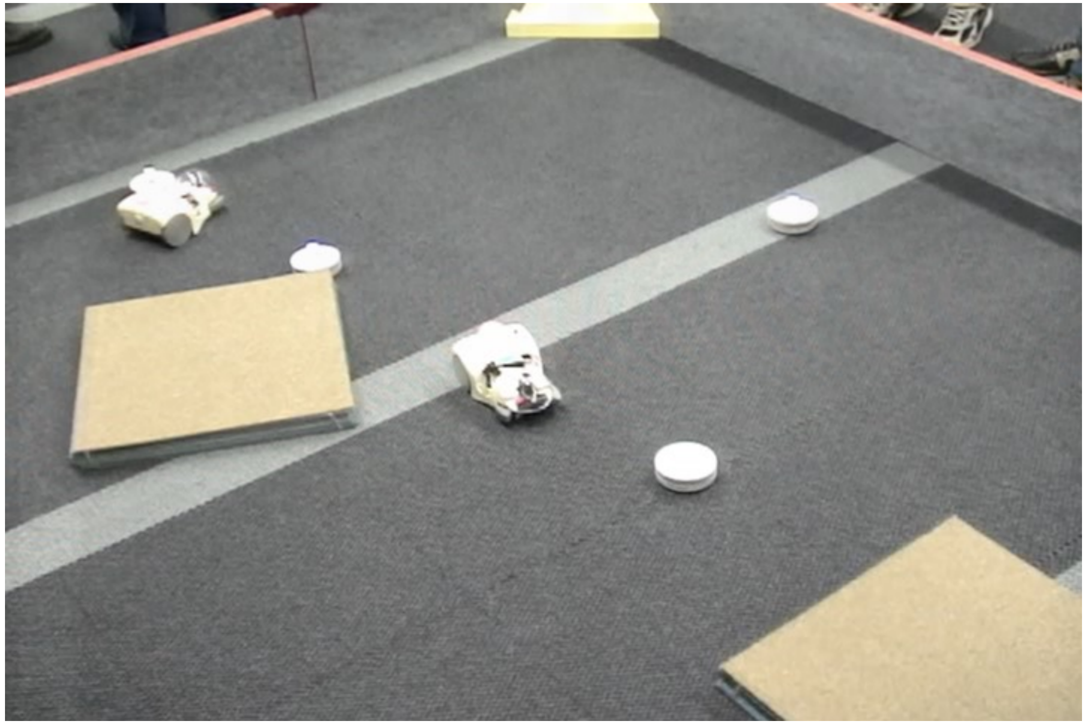


Learning to Survive and Reproduce

- Catch battery packs
  - survival



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

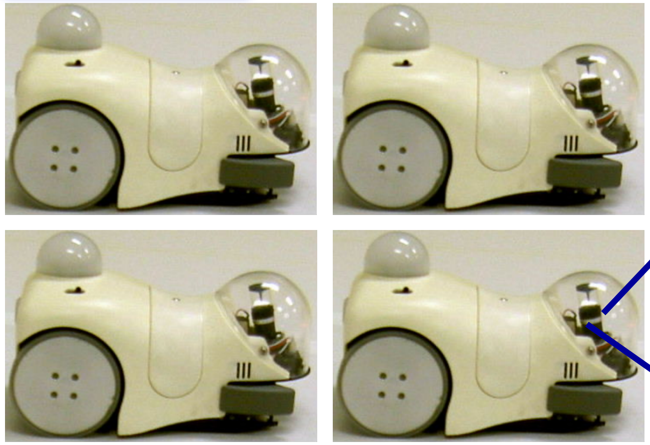


(Doya & Uchibe, 2005)

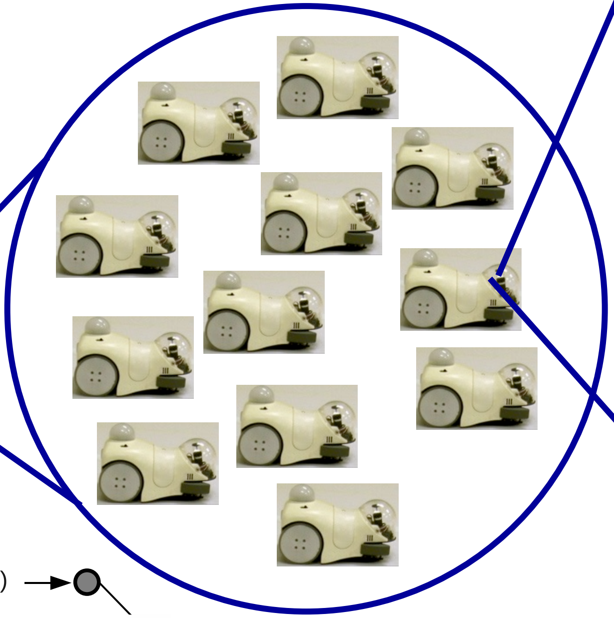
Embodied Evolution (Elfwing et al., 2011)

Population

Robots



Virtual agents 15-25



Genes

Weights for top layer NN

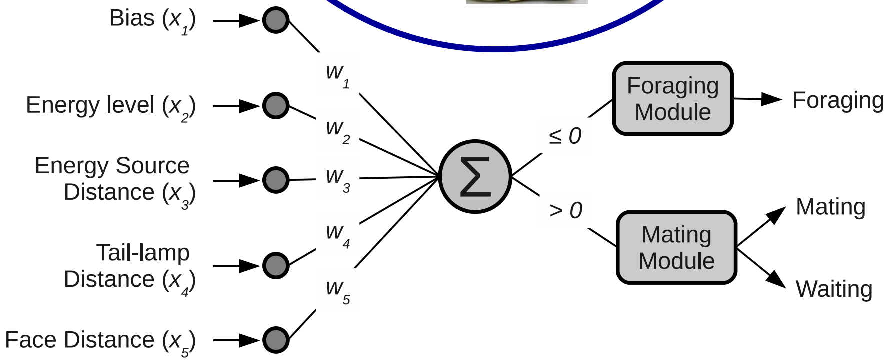
$w_1, w_2, \ldots, w_n$

Weights shaping rewards

$v_1, v_2, \ldots, v_n$

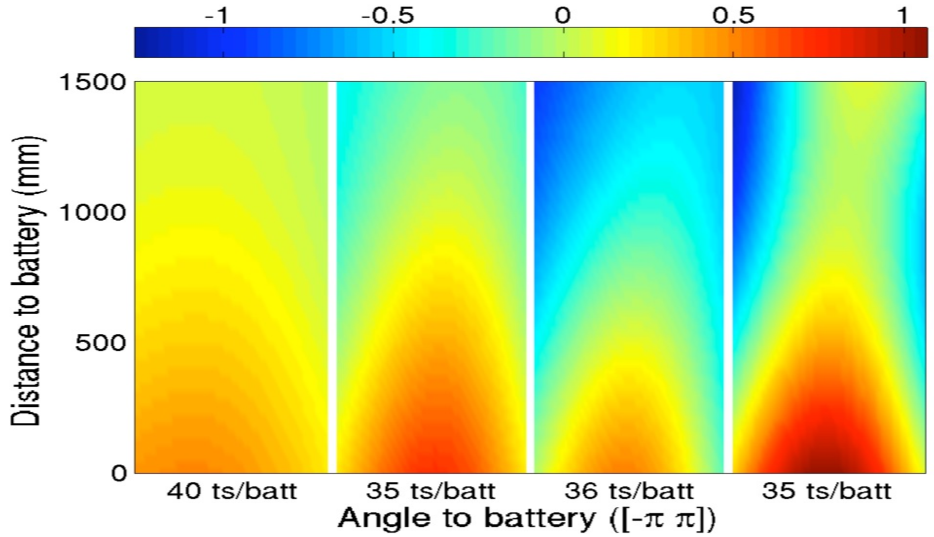
Meta-parameters

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

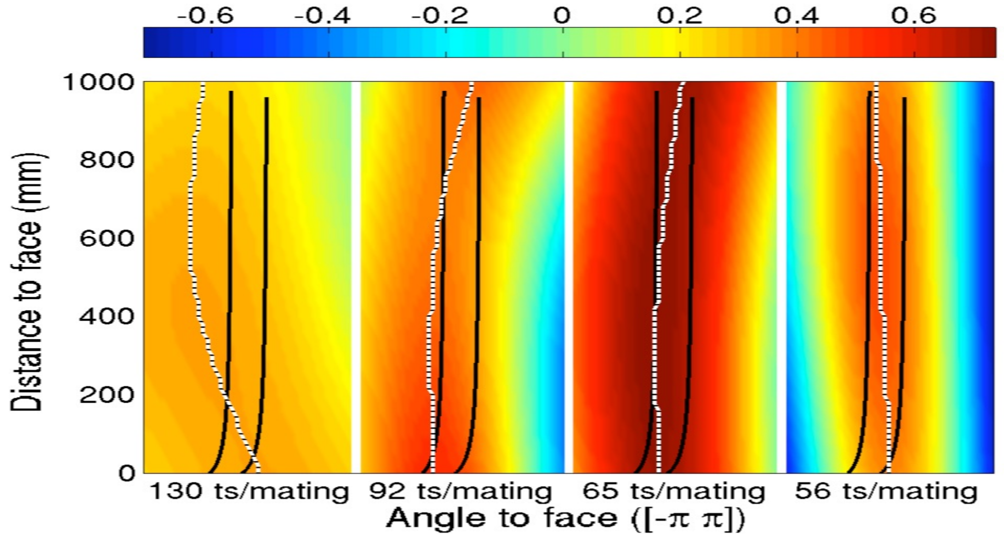


Evolution of Shaping Rewards

- Vision of battery



- Vision of face

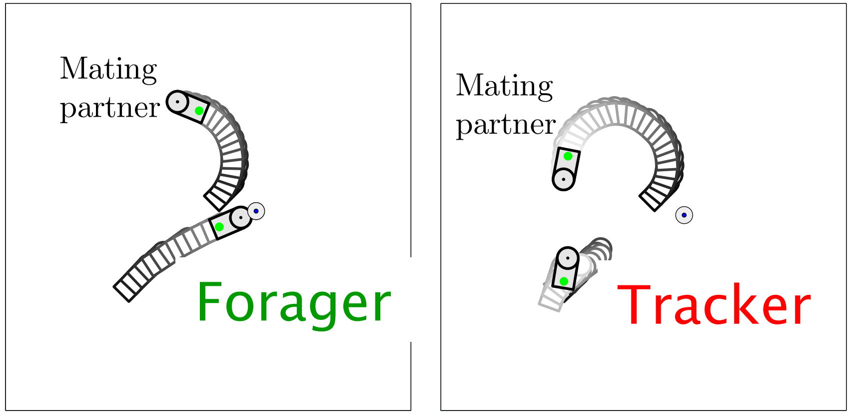


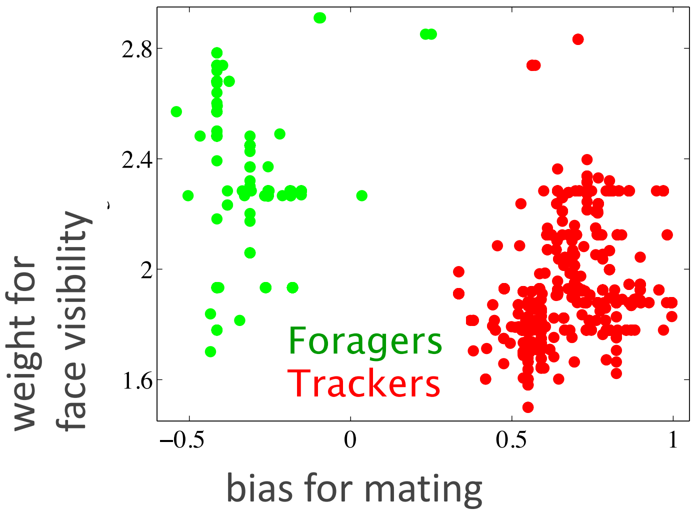
(Elfwing et al., 2011)

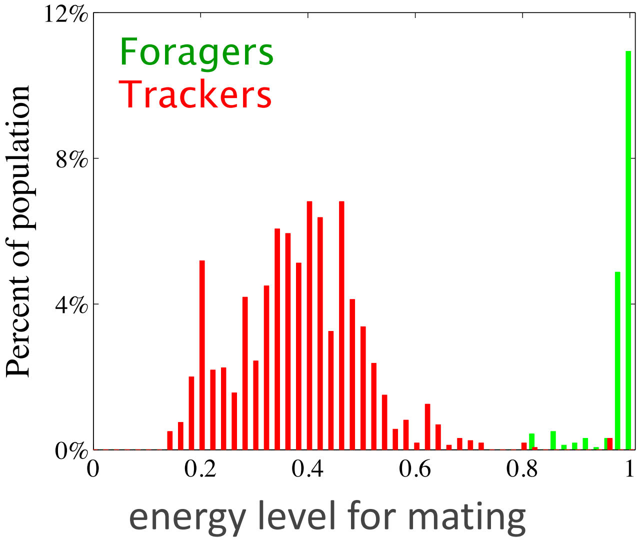
Polymorphism within Colony

(Elfwing et al. 2014)

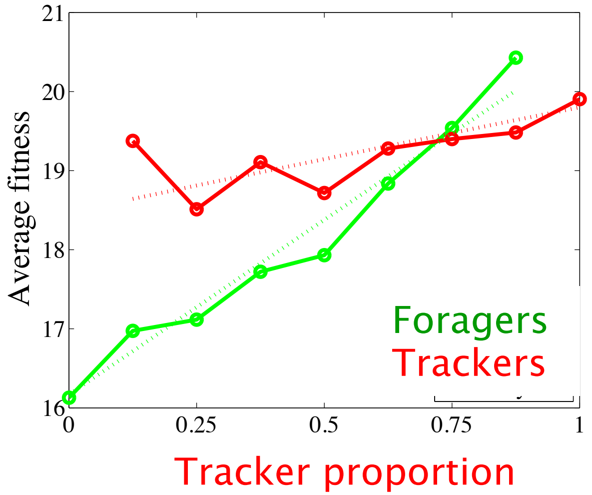
- Foragers and Trackers





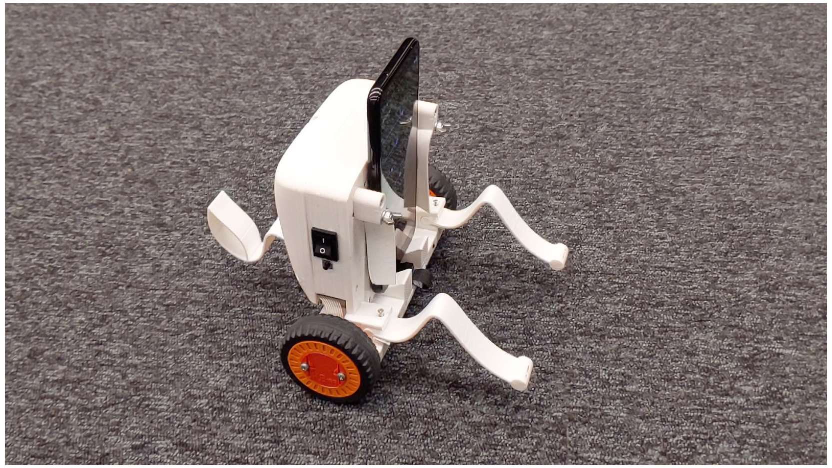


- Evolutional stability

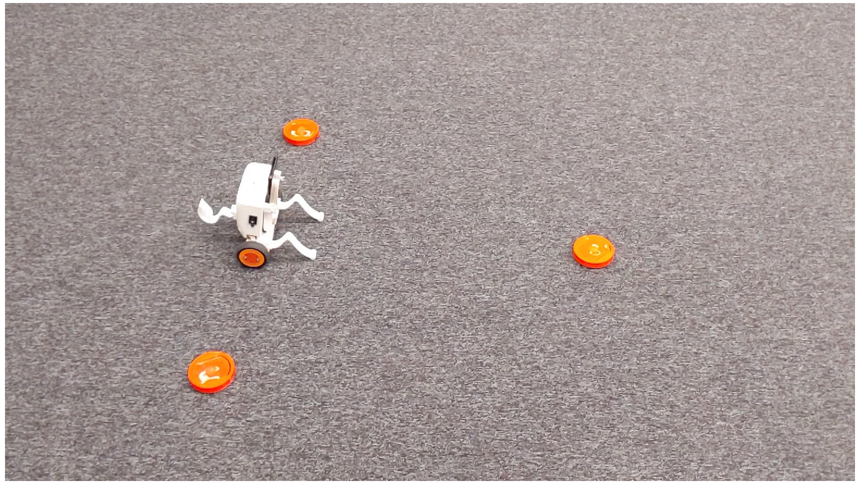


Smartphone Robot Project

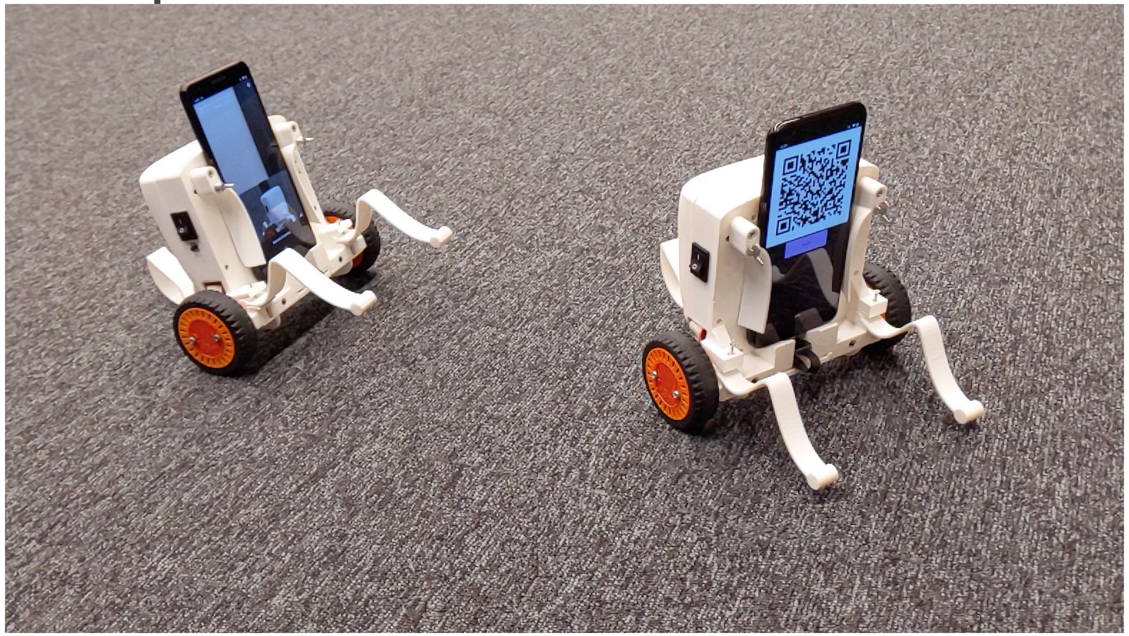
- Motor control



- Survival



- Reproduction



- Evolution of extrinsic/intrinsic rewards
- Meta-learning
- Acquisition of internal models

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

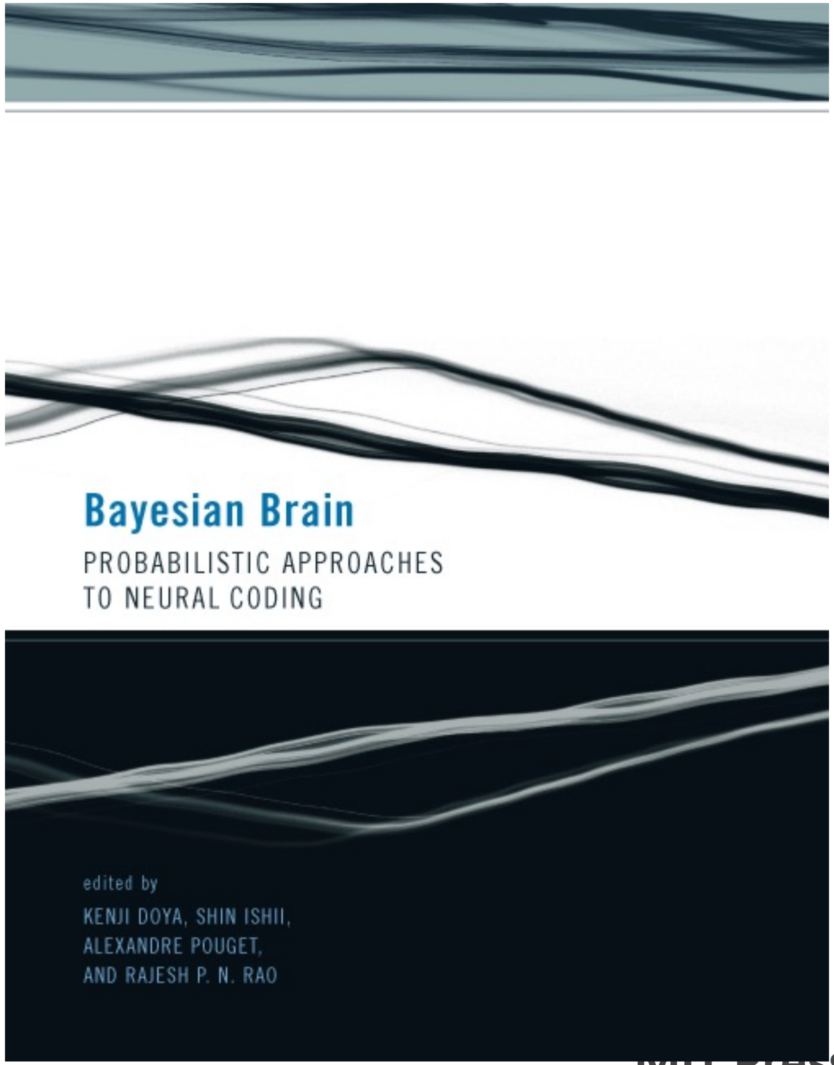
P(X)
0.5
0.3
0.2

- 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

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

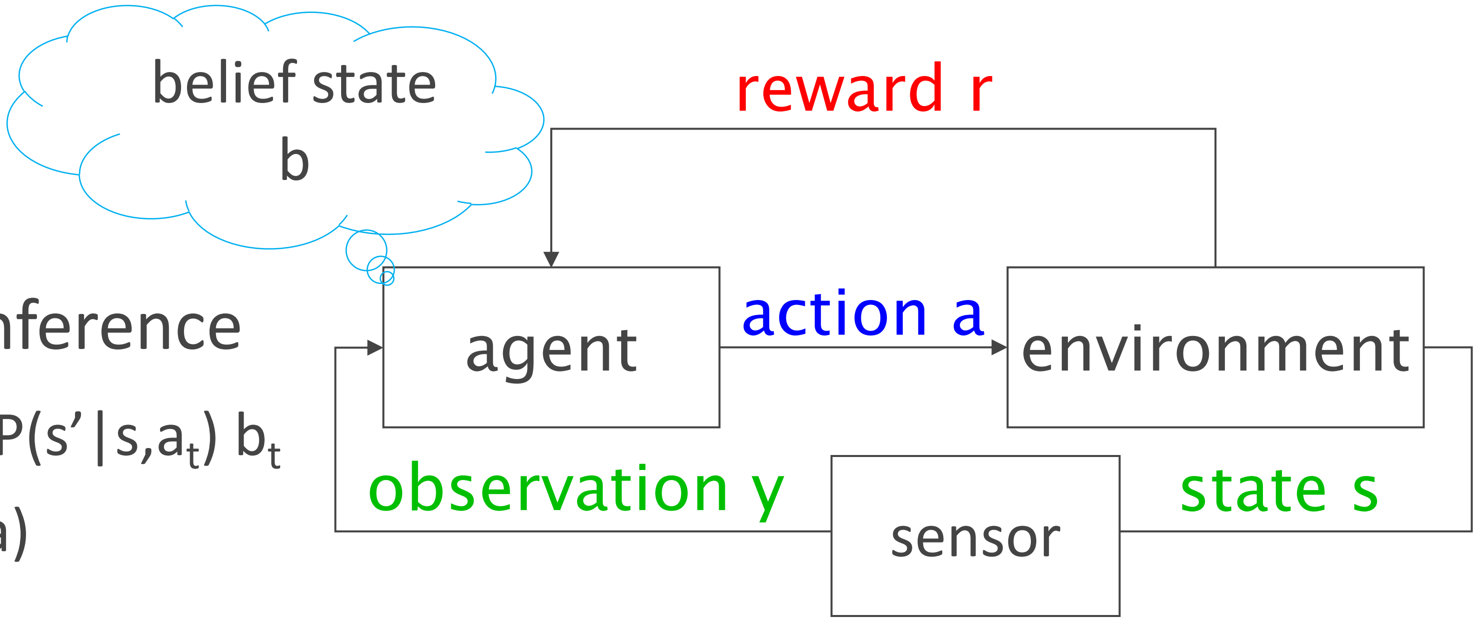
Partially Observable Markov Decision Process (POMDP)

- 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

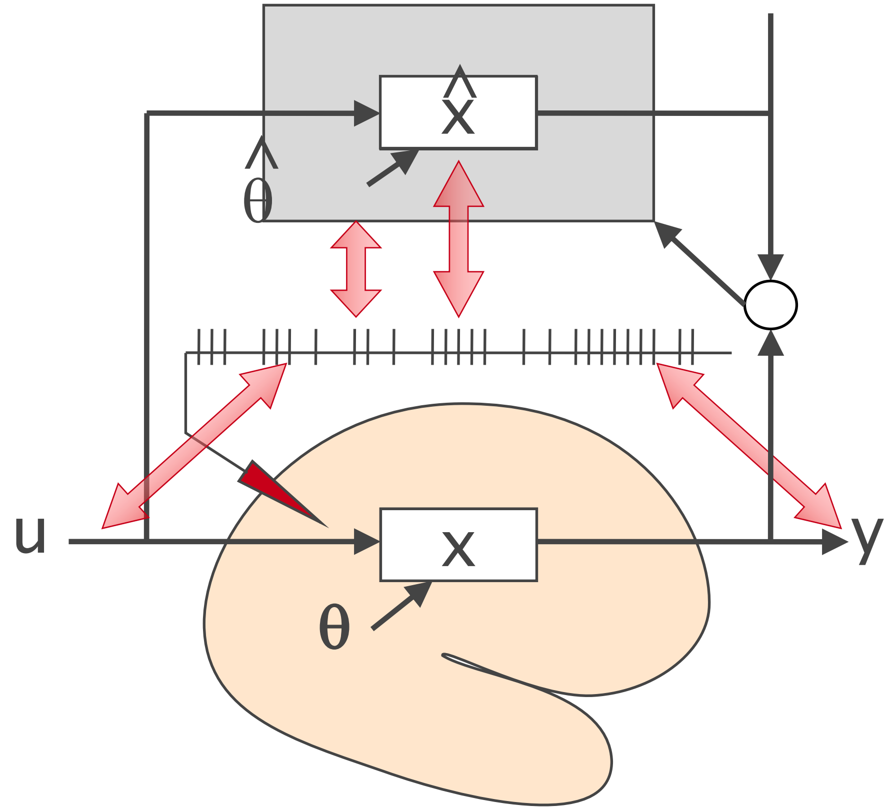
  $b_{t+1} \propto P(y_{t+1} | s') \sum_s P(s' | s, a_t)\ b_t$

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



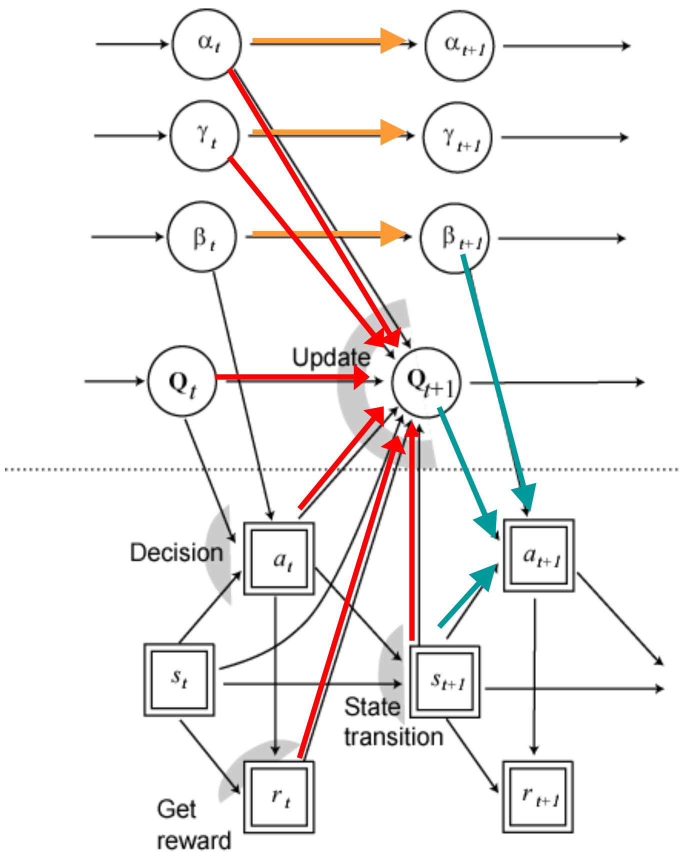
Model-based Neural Analysis

- 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



Bayesian Inference of Action Values (Samejima et al. 2004)

- Hidden variables
  - $x=(Q,\alpha,\beta,\gamma)$
  - $p(x'|x)$: learning rule
- Observable variables
  - $y=(s,a,r)$
  - $p(y|x)$: action policy
- Predictive prior
  - $P(x_{t+1}|y_{1:t}) = \int P(x_{t+1}|x_t)P(x_t|y_{1:t})dx_t$
- Posterior given observation y_{t+1}
  - $P(x_{t+1}|y_{1:t+1}) \propto P(y_{t+1}|x_{t+1})P(x_{t+1}|y_{1:t})$



Opinion TRENDS in Neurosciences Vol.27 No.12 December 2004 Full text

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

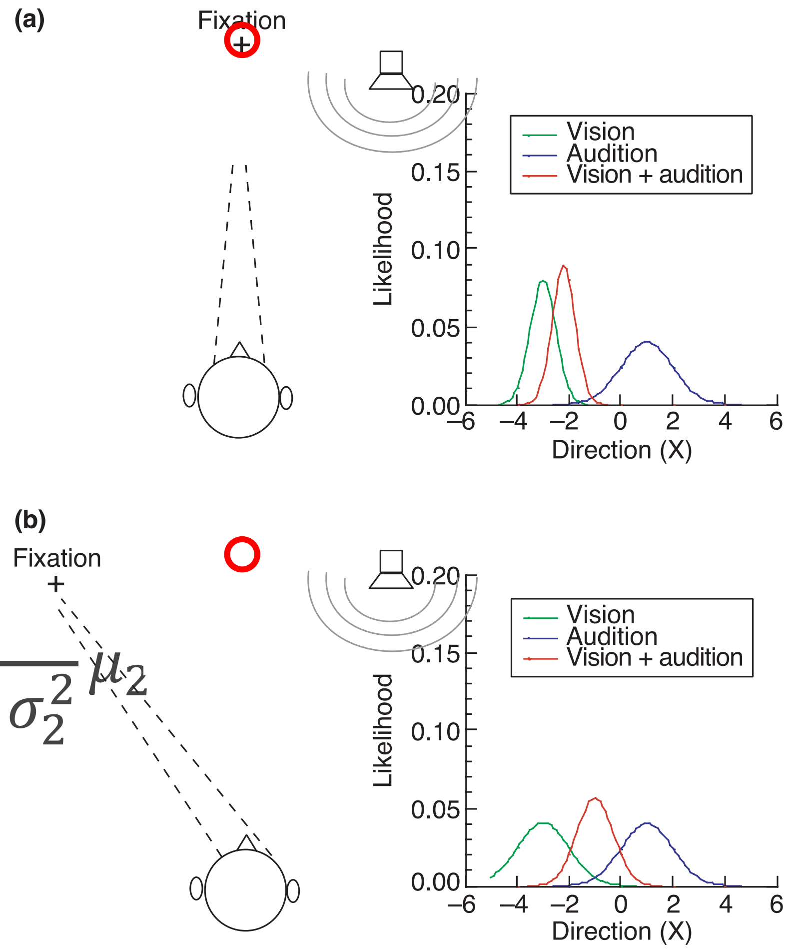
David C. Knill and Alexandre Pouget

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

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

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

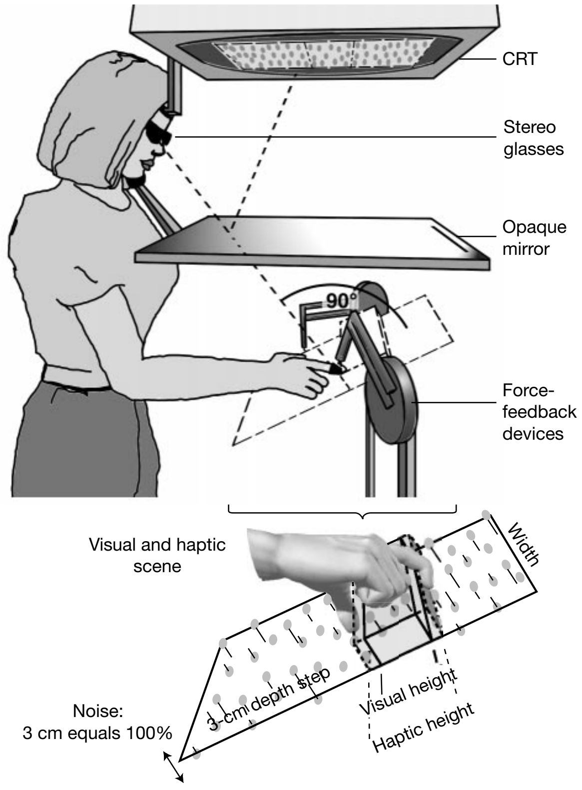
$$\sigma^2 = \frac{\sigma_1^2\sigma_2^2}{\sigma_1^2+\sigma_2^2}$$

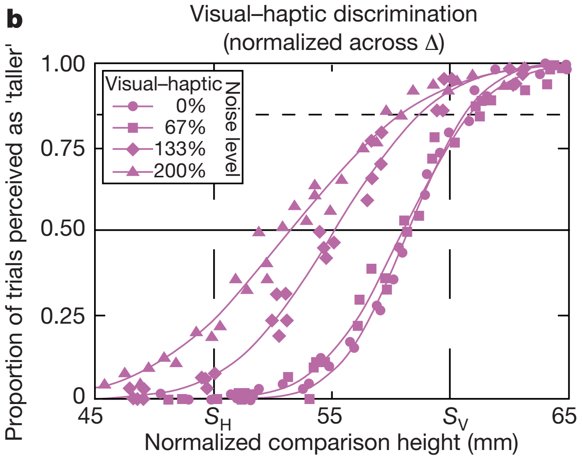


Multi-Sensory Integration

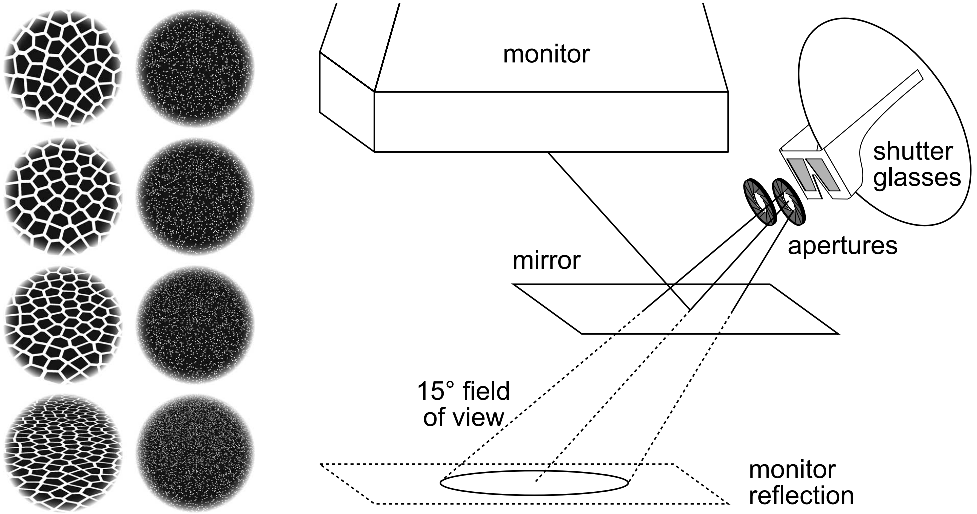
Humans integrate visual and haptic information in a statistically optimal fashion (2002, Nature)

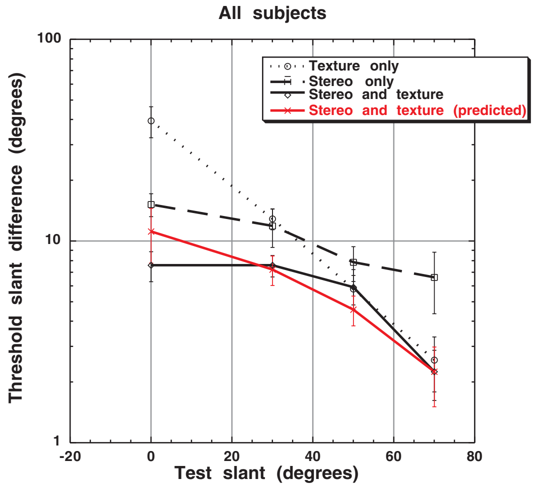
Marc O. Ernst* & Martin S. Banks





Knill & Saunders, (2003, Vision Research)

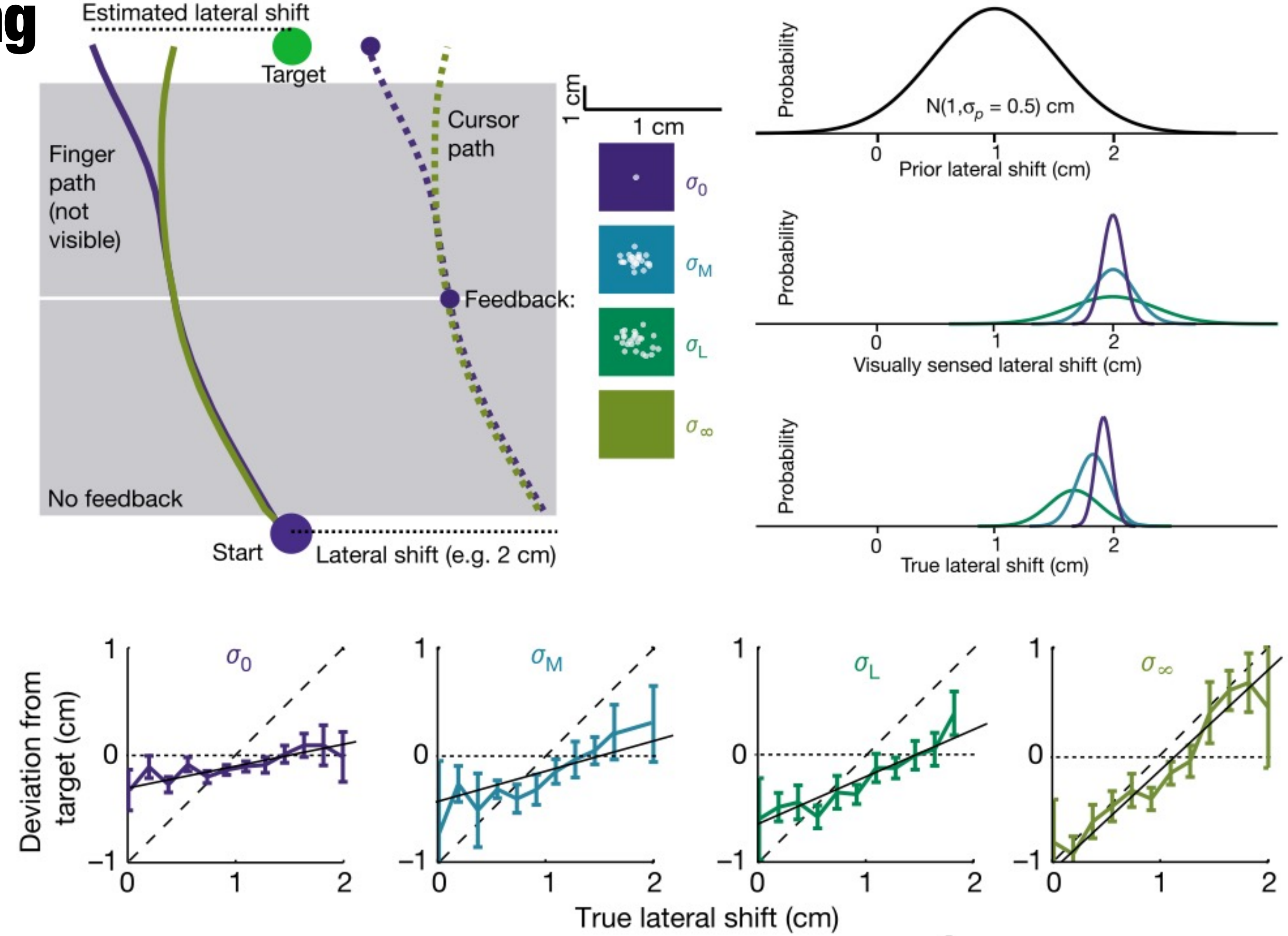




letters to nature (2004)

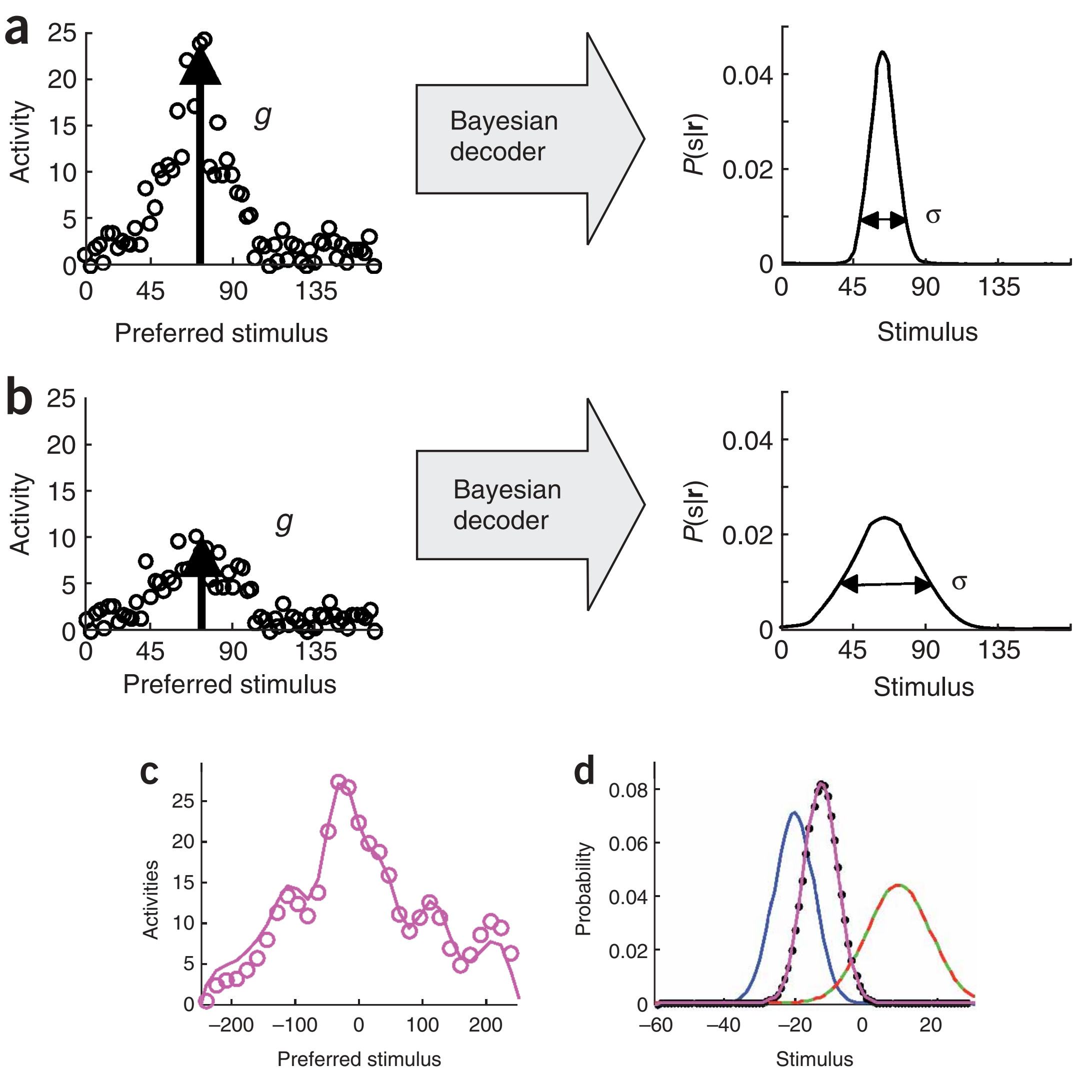
Bayesian integration in sensorimotor learning

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)

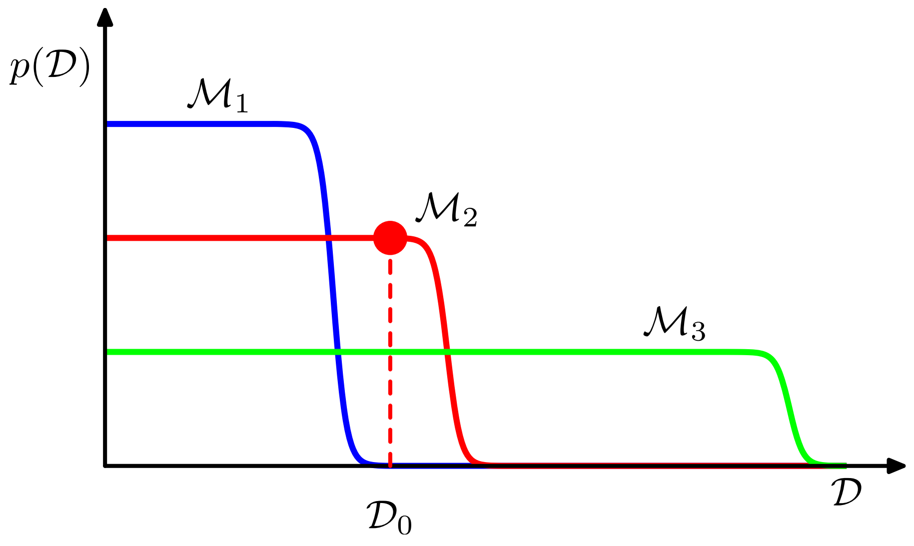


Bayesian Model Selection

- Bayes rule: $P(\theta|Y) = P(Y|\theta)\, P(\theta) / P(Y)$
- Denominator: marginal likelihood

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

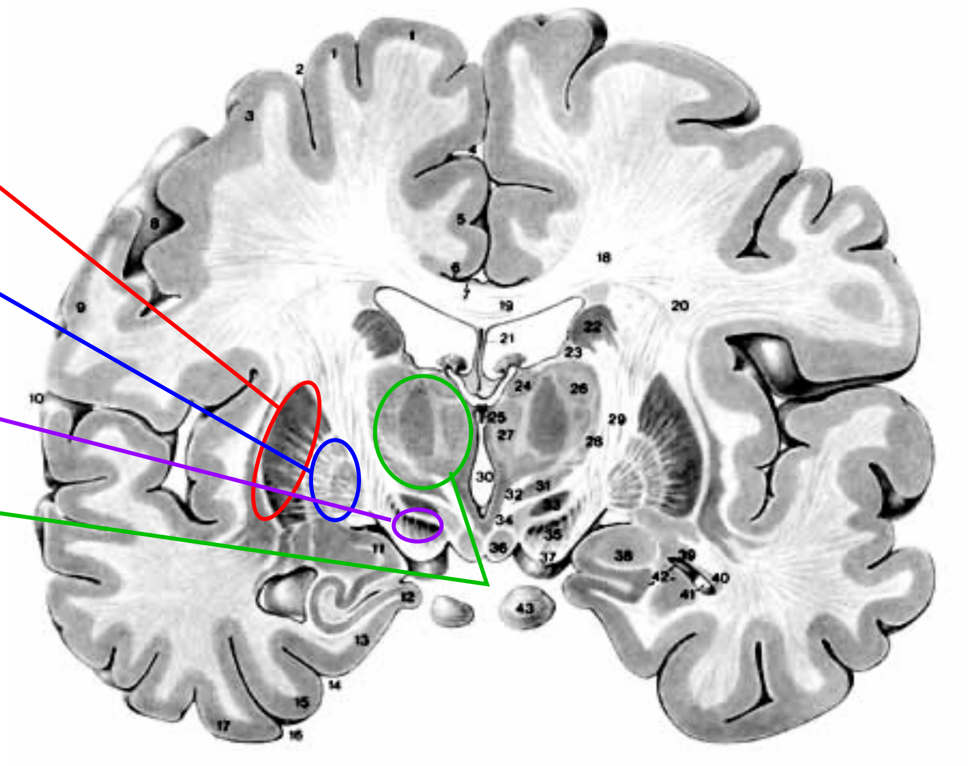
  - Measure of compatibility of model and data
- Too simple model
  - likelihood $P(Y|\theta)$ is low
- Too complex model
  - penalized by thin $P(\theta)$
- 'Evidence' of model

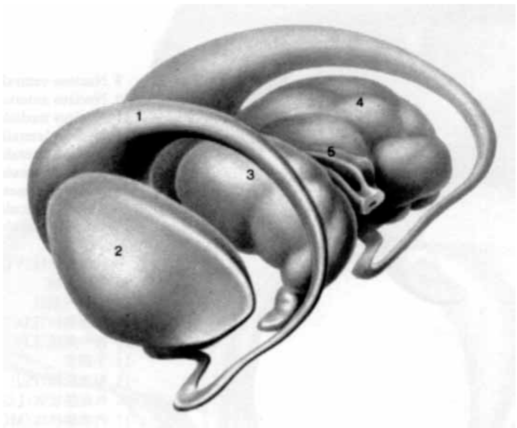


Basal Ganglia

- Locus of Parkinson’s and Huntington’s diseases

Striatum
Globus Pallidus
Substantia Nigra
Thalamus

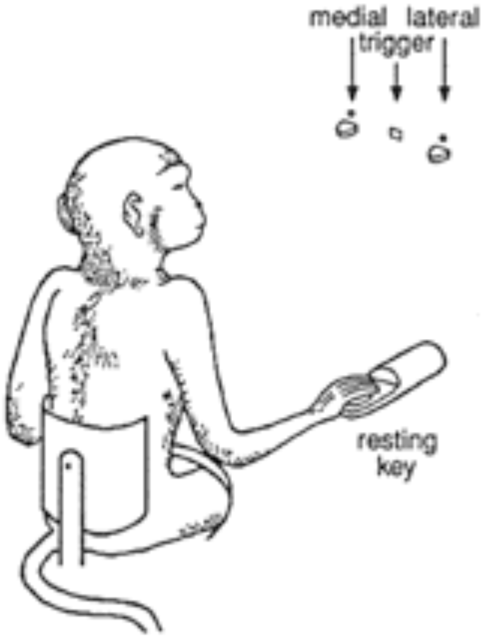


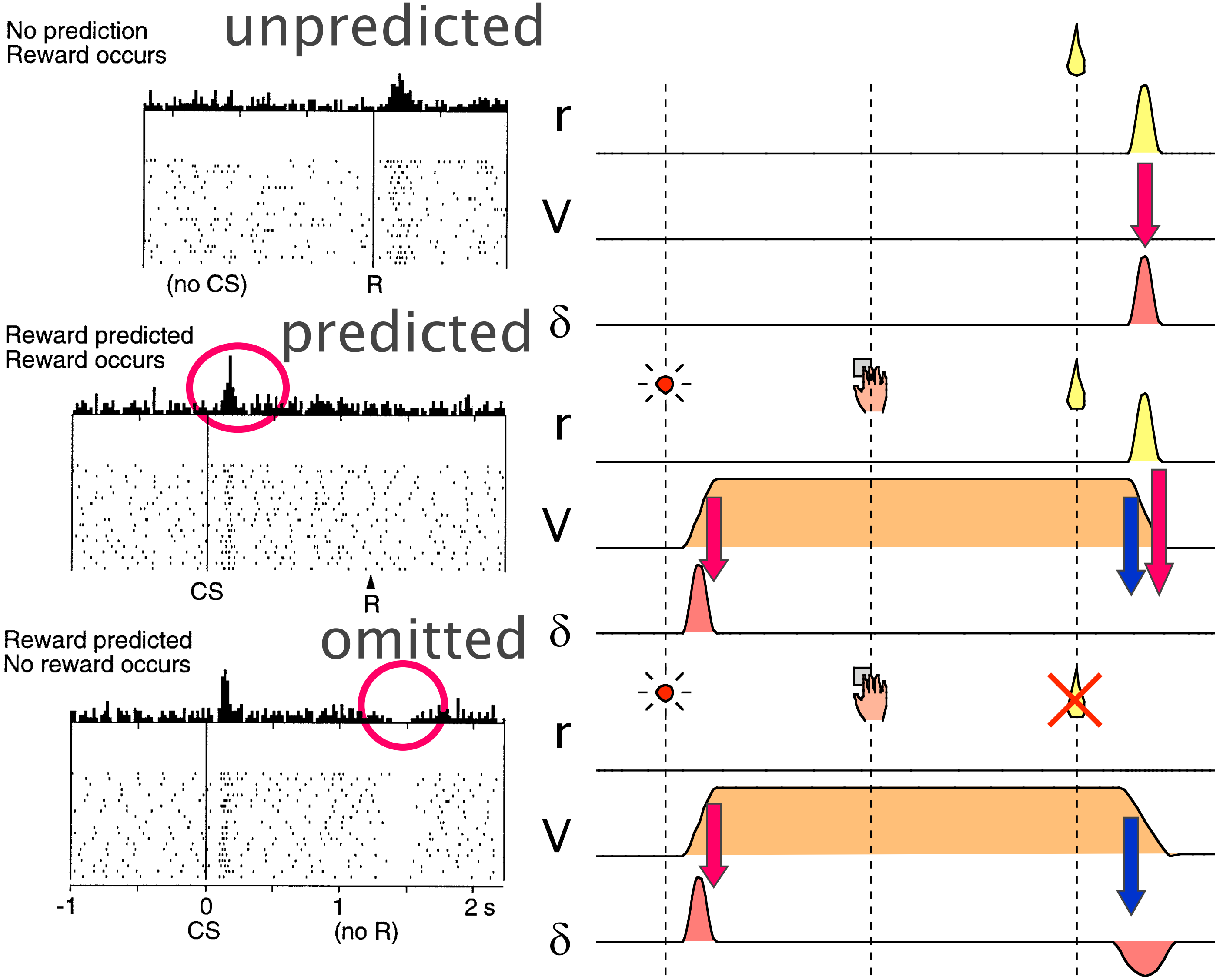


- What is their normal function??

Dopamine Neurons Code TD Error

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

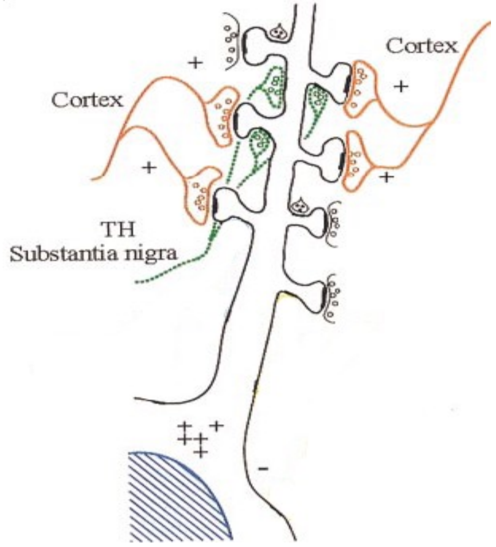


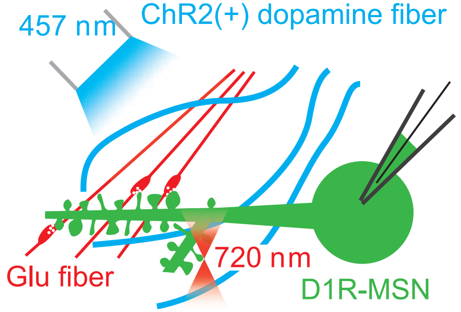


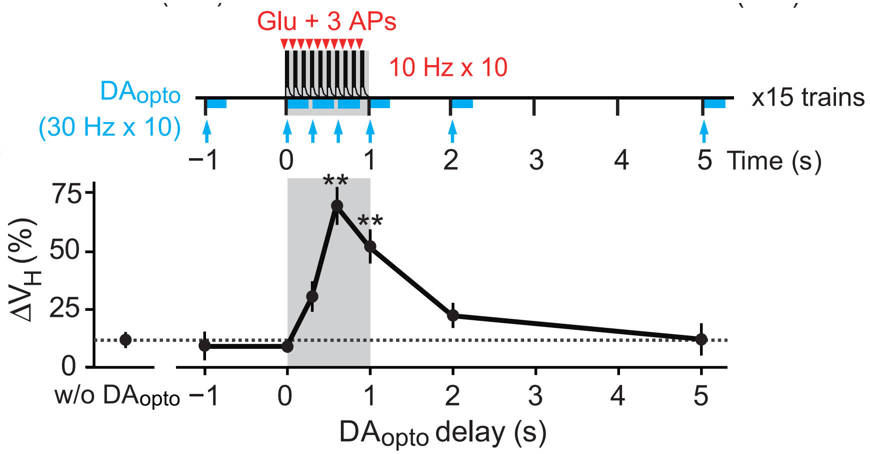
(Schultz et al. 1997)

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)

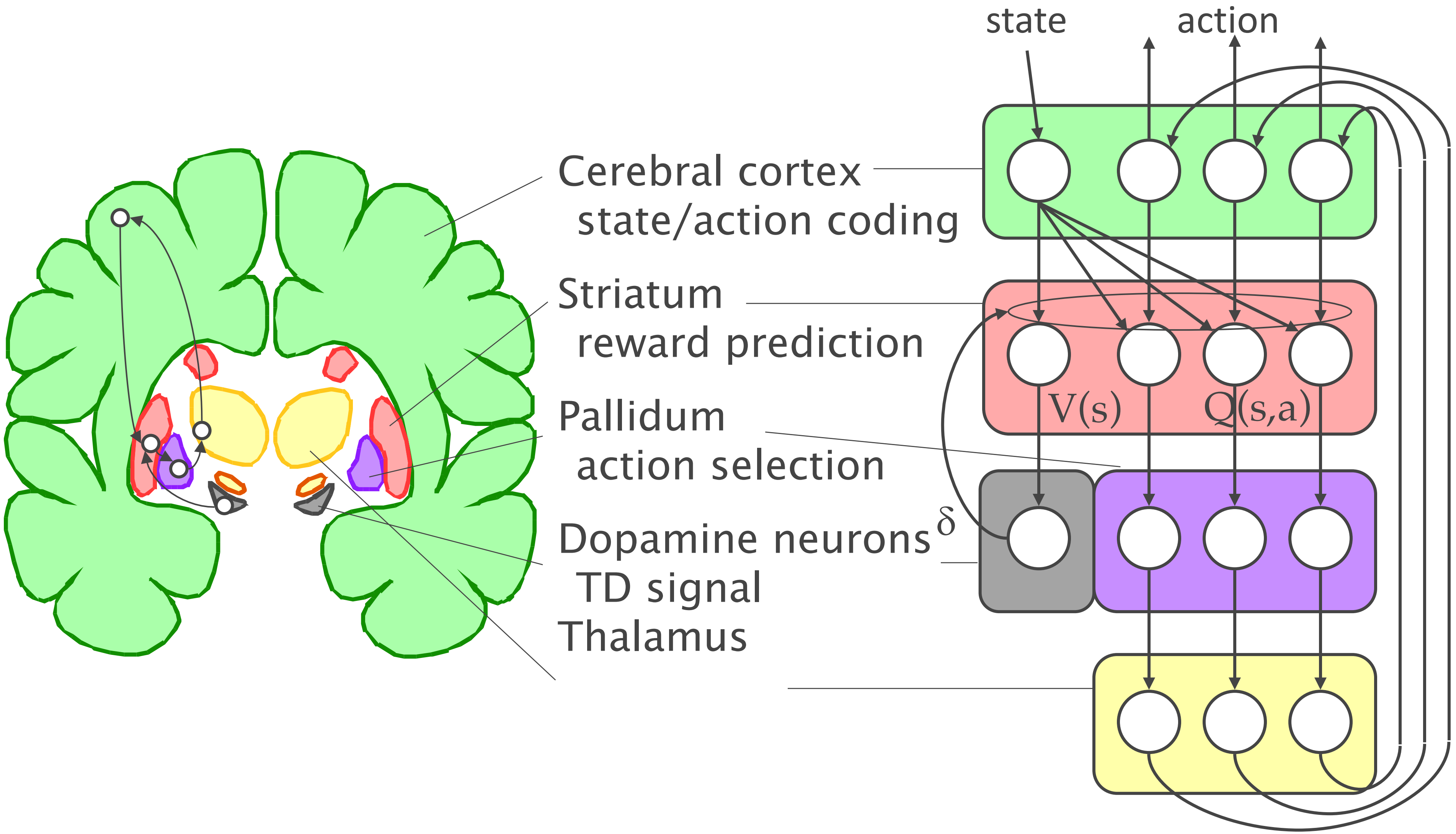






Basal Ganglia for Reinforcement Learning?

(Doya 2000, 2007)

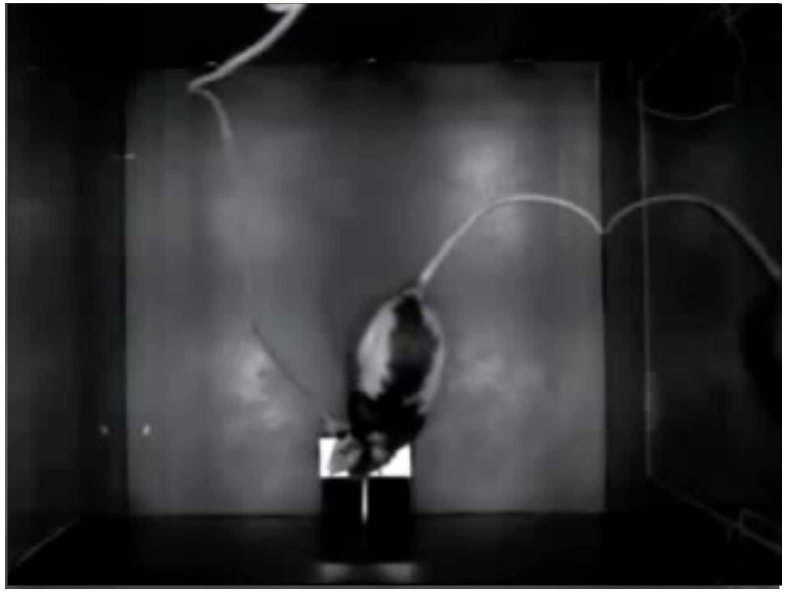


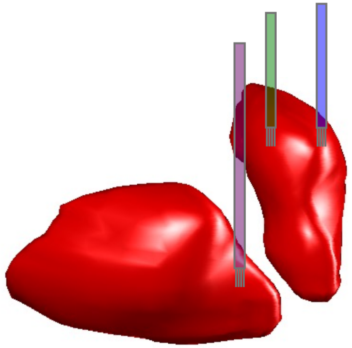
Distinct Neural Representation in the Dorsolateral, Dorsomedial, and Ventral Parts of the Striatum during Fixed- and Free-Choice Tasks

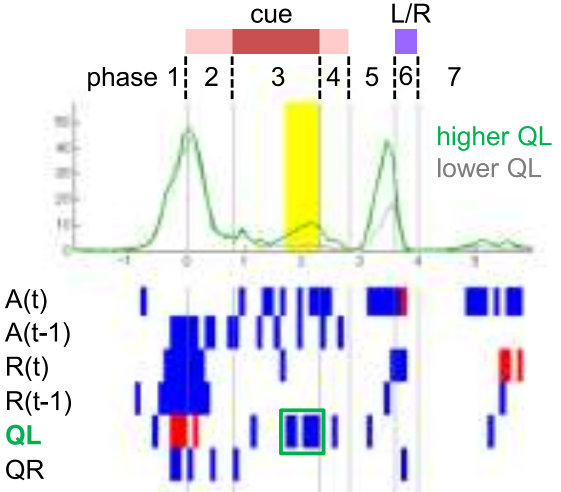
Makoto Ito and Kenji Doya

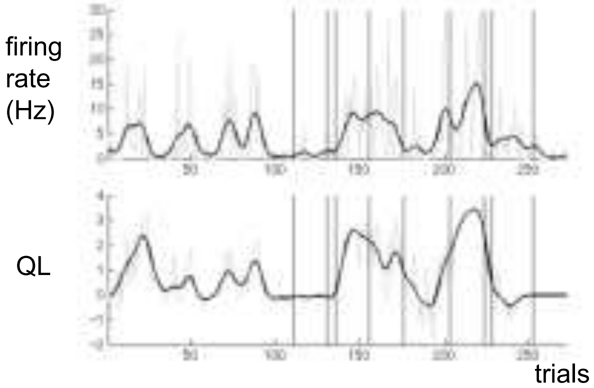
The Journal of Neuroscience, 2015

Left Center Right

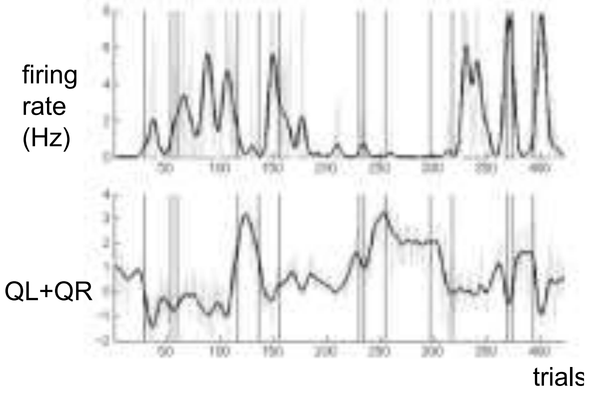








- Dorsolateral
  - movements
- Dorsomedial
  - action value
- Ventral
  - state value



Generalized Q-learning Model

(Ito & Doya, 2009)

- Action selection

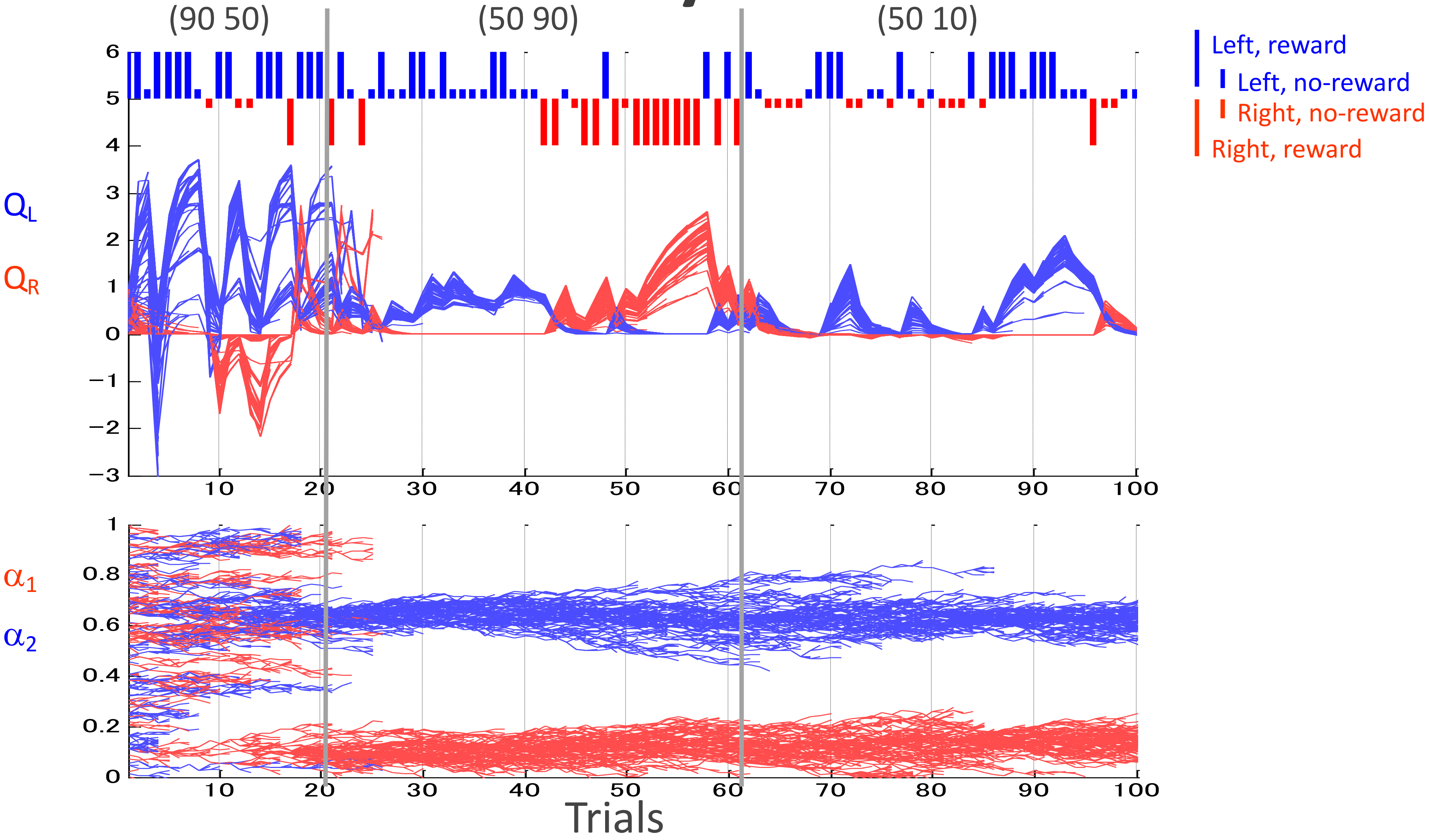
  $P(a(t)=L) = \exp Q_L(t)/(\exp Q_L(t)+\exp Q_R(t))$

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

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

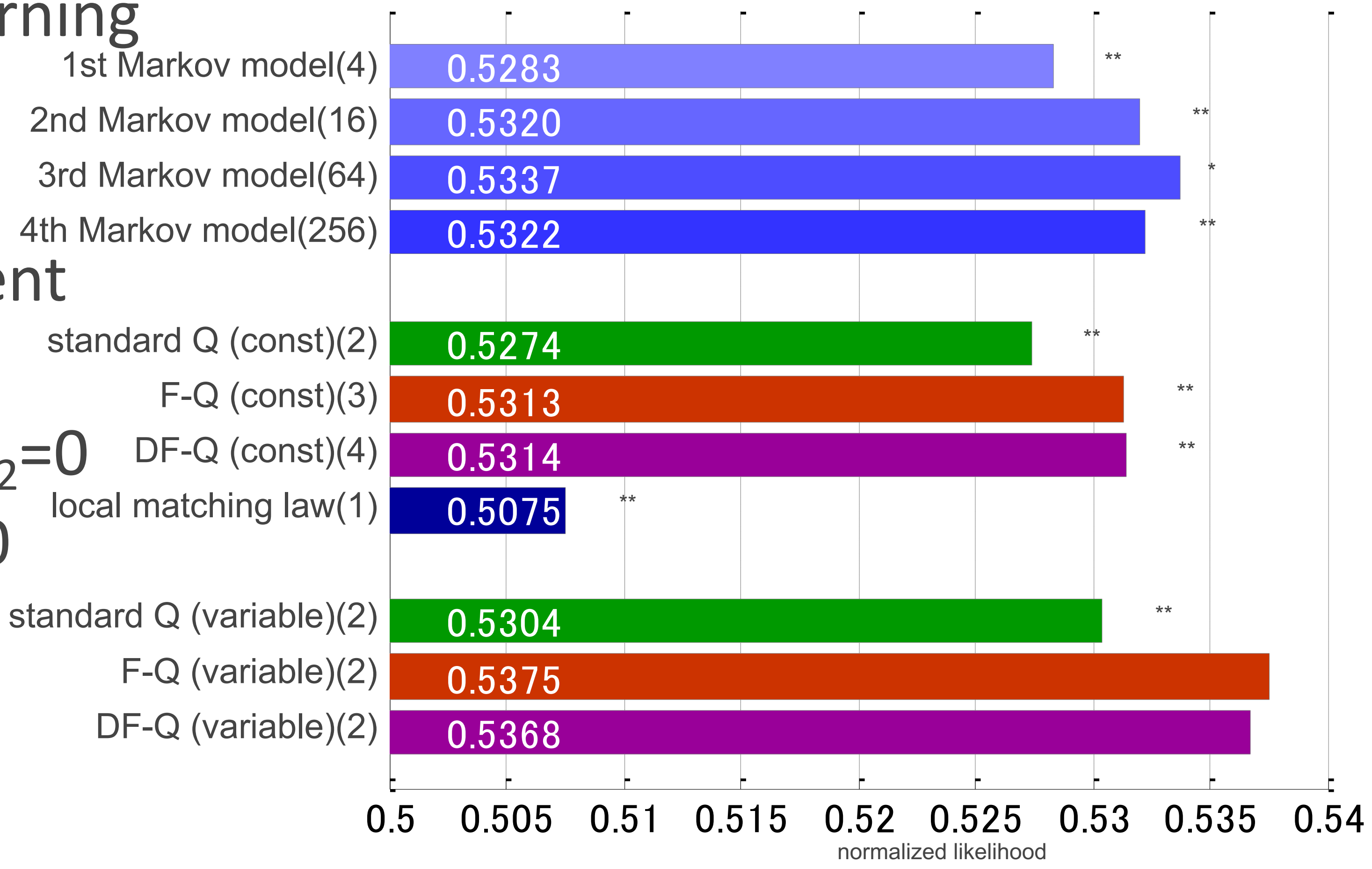
- Parameters
  - α_1: learning rate
  - α_2: forgetting rate
  - κ_1: reward reinforcement
  - κ_2: no-reward aversion

Estimation by Particle Filter

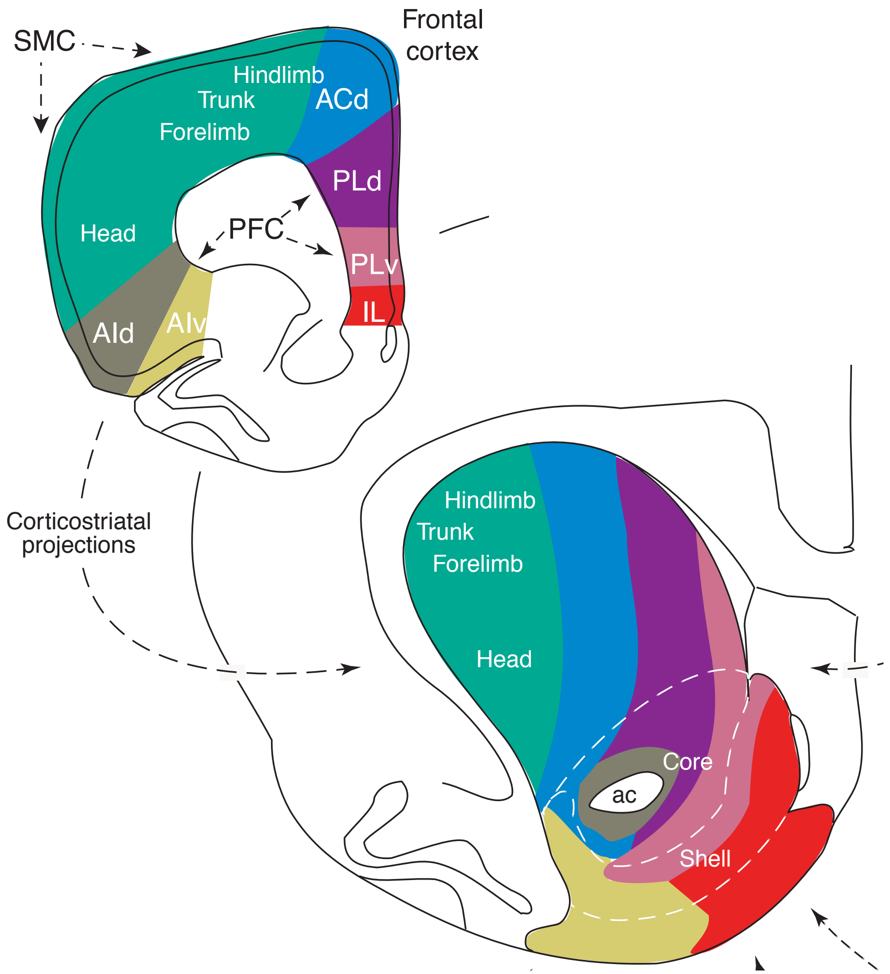


Model Fitting

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



Hierarchy in Cortico-Striatal Network

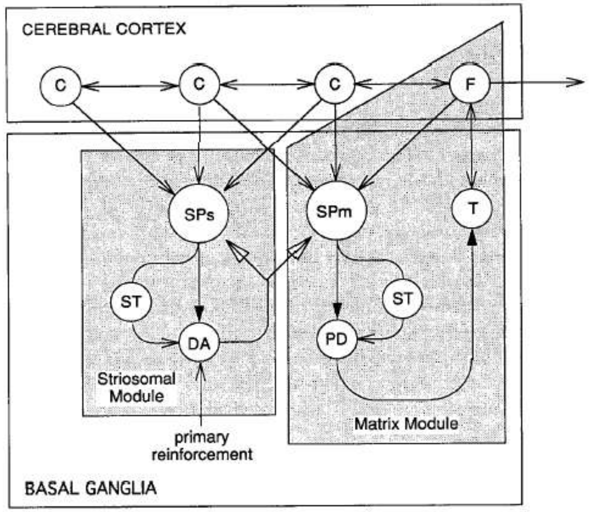


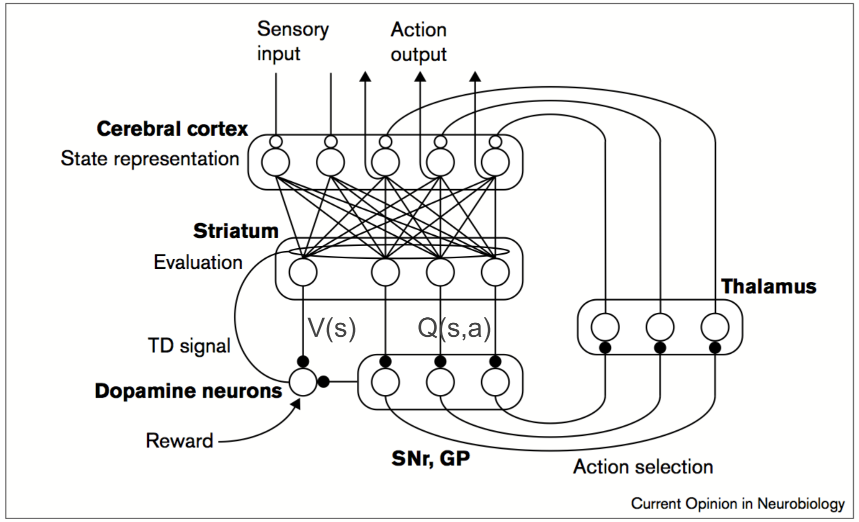
(Voorn et al., 2004)

- 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?

Striosome Neurons as Critic?

- Actor-critic (Houk et al., 1995) or state/action value (Doya, 2000)





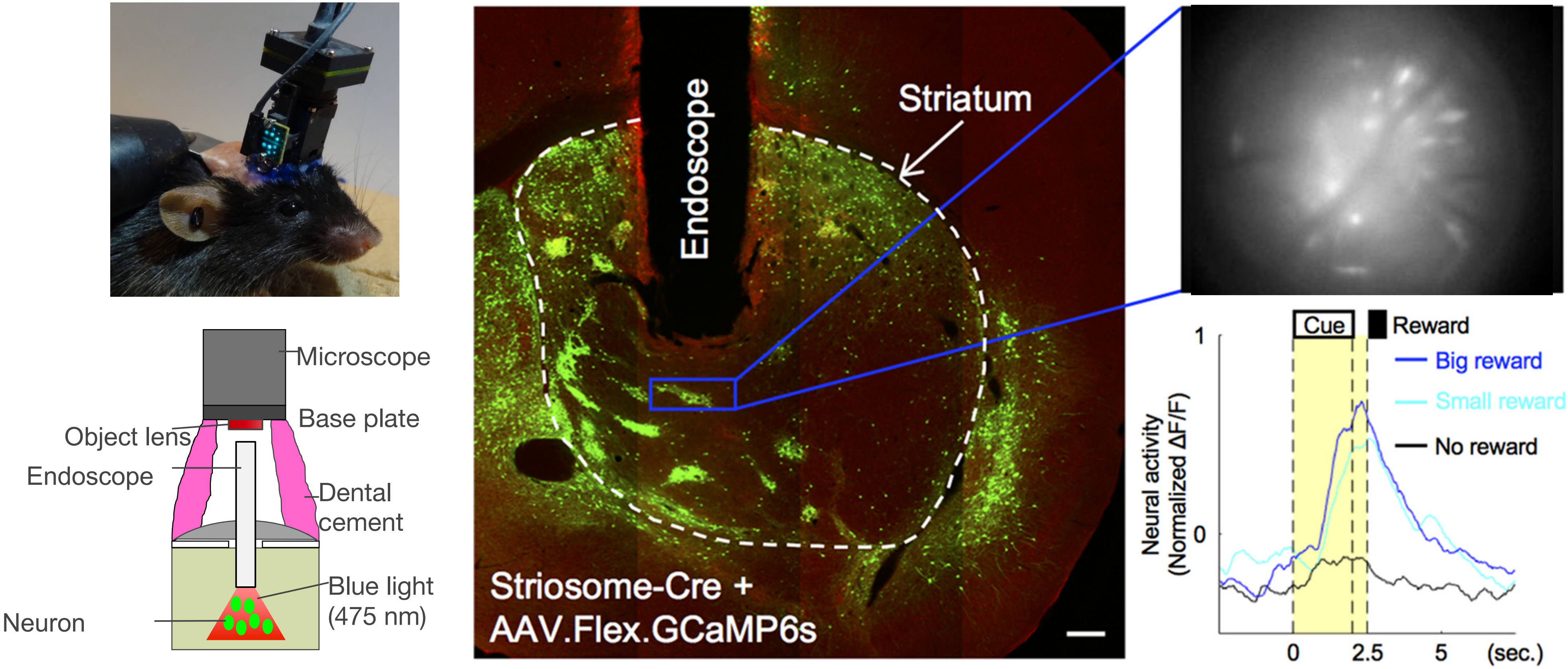
  - Do striosome neurons code state value?
  - Do matrix neurons code action or action value?
- Need cell-type specific recording
  - optolodes or calcium imaging

eNeuro (2018)

Reward-Predictive Neural Activities in Striatal Striosome Compartments

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

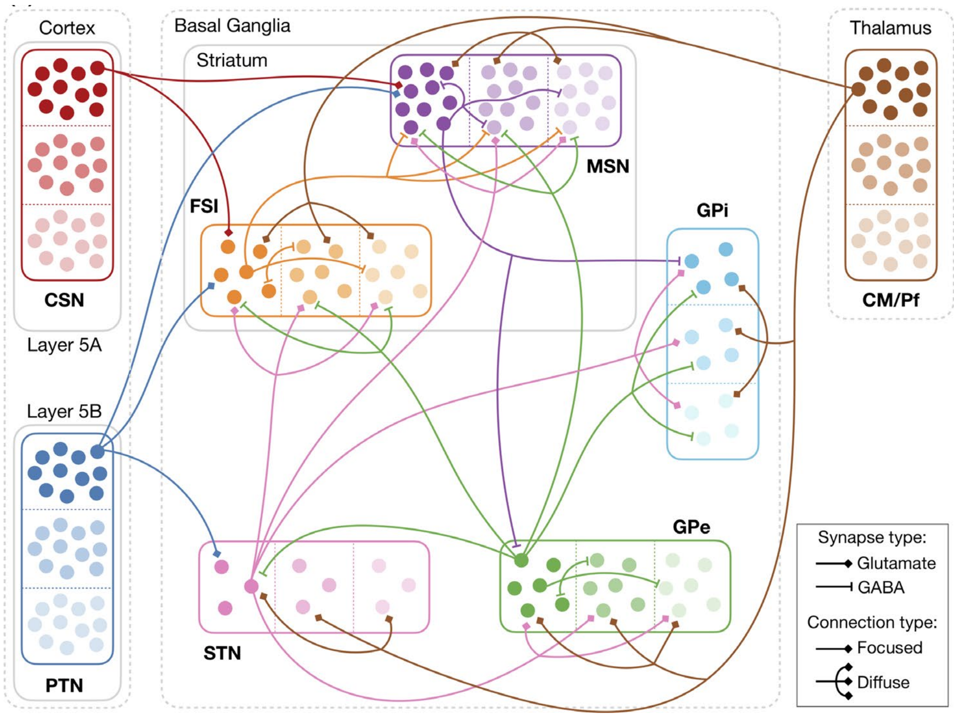
Imaging striosome neuron activity by endoscope



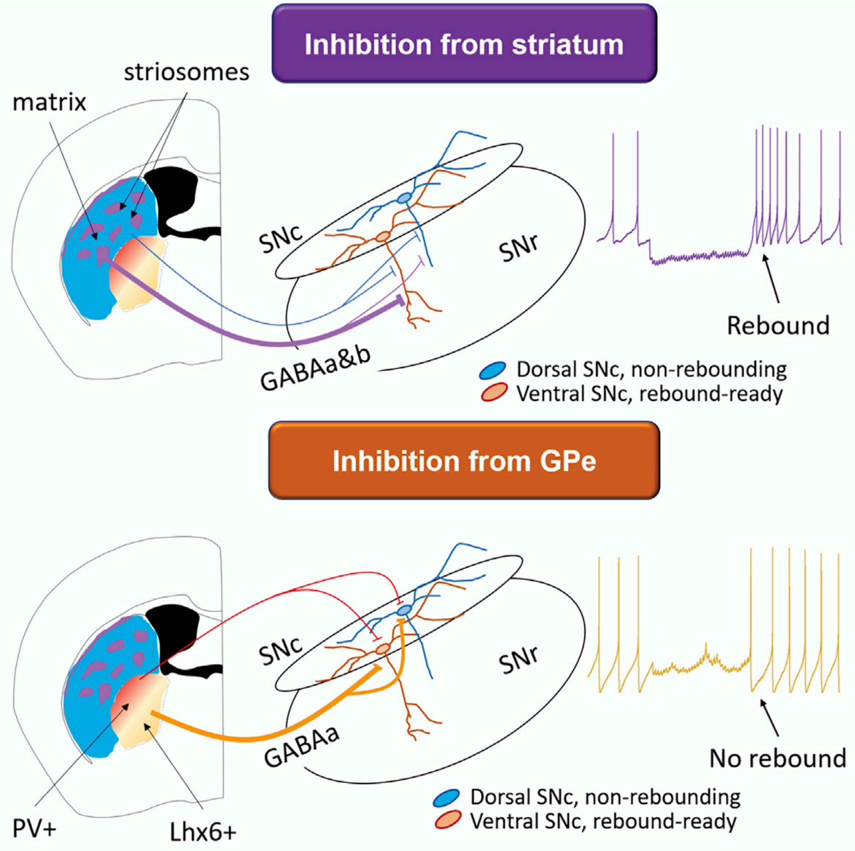
Open Questions about Basal Ganglia

Parallel, multi-inhibitory pathways

(Girard et al. 2020)



TD like response of dopamine neurons



(Evans et al. 2020)

Model-free and Model-based RL

Model-free RL

- Memorize action values
  - Q(state, action)
- Reactive action
  - $P(a|s) \sim \exp[\beta Q(s,a)]$
- On-line learning by TD error
  - $\delta = \text{reward} + \gamma Q(s',a') - Q(s,a)$

Simple, but slow learning

Model-based RL

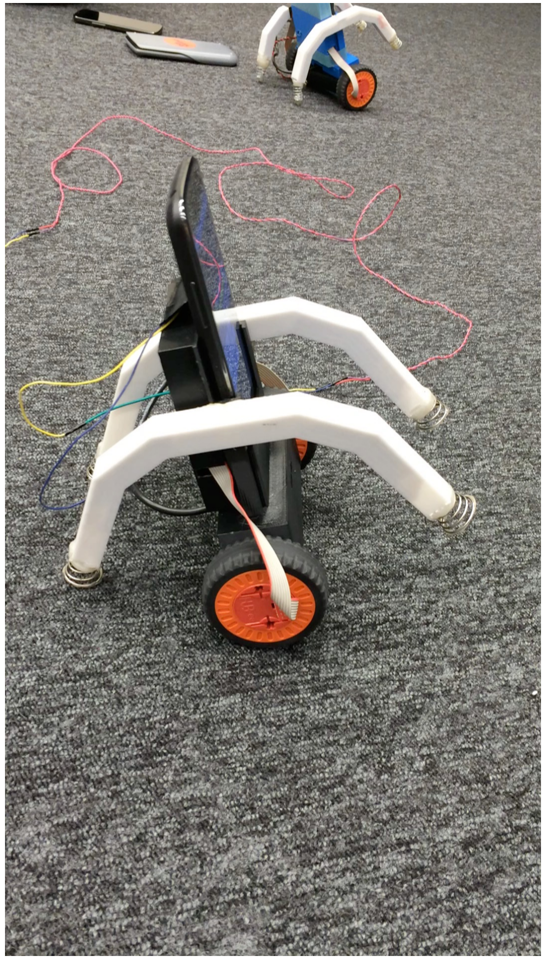
- Learn internal models
  - P(next state | state, action)
  - R(state, action)
- Estimate current state
  - $P(s_t|o_t,a_{t-1}) \propto P(o_t|s_t)\Sigma_{s_{t-1}} P(s_t|s_{t-1},a_{t-1})P(s_{t-1})$
- Predict values
  - $Q(s,a) = \Sigma_{s'} P(s'|s,a)[R(s,a)+\gamma V(s')]$
  - $V(s)=\max_a \Sigma_{s'} P(s'|s,a)[R(s,a)+\gamma V(s')]$

Flexible, but heavy load

Bounce Up and Balance by PILCO

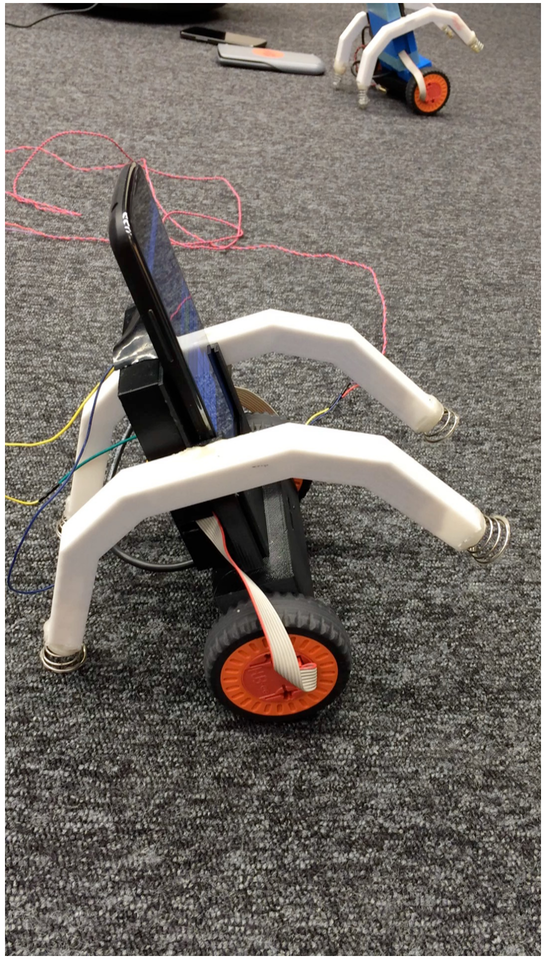
(Paavo Parmas)

1st try



2nd try

8th try



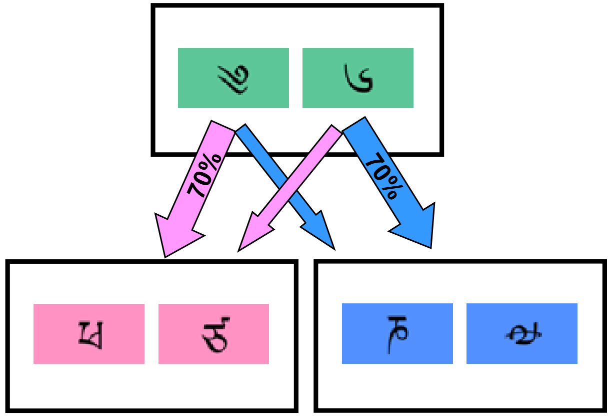
Mental Simulation

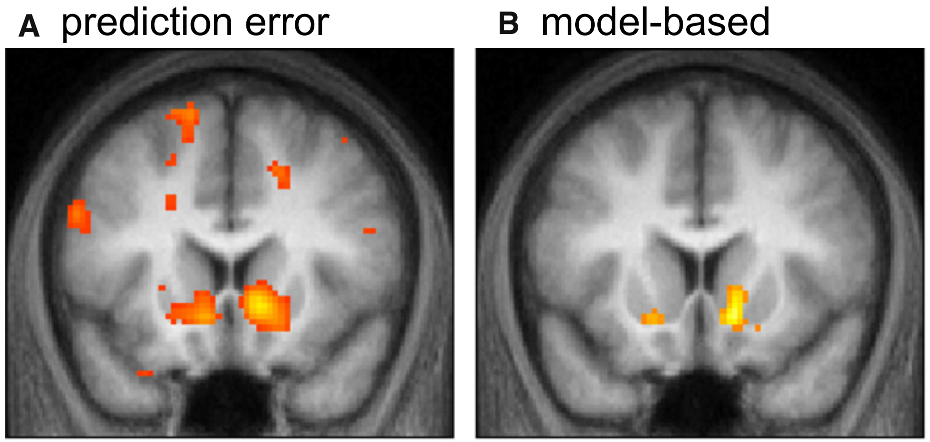
Brain's process using
an action-dependent state transition model
$s'=f(s,a)$ or $P(s'|s,a)$

- 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,…

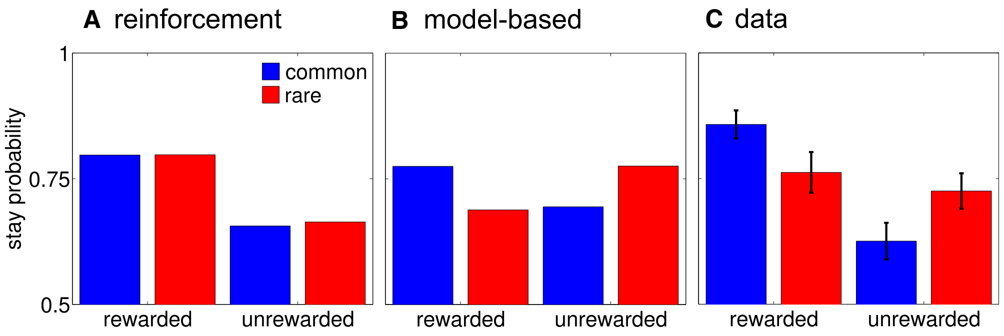
Model-free and Model-based Choice

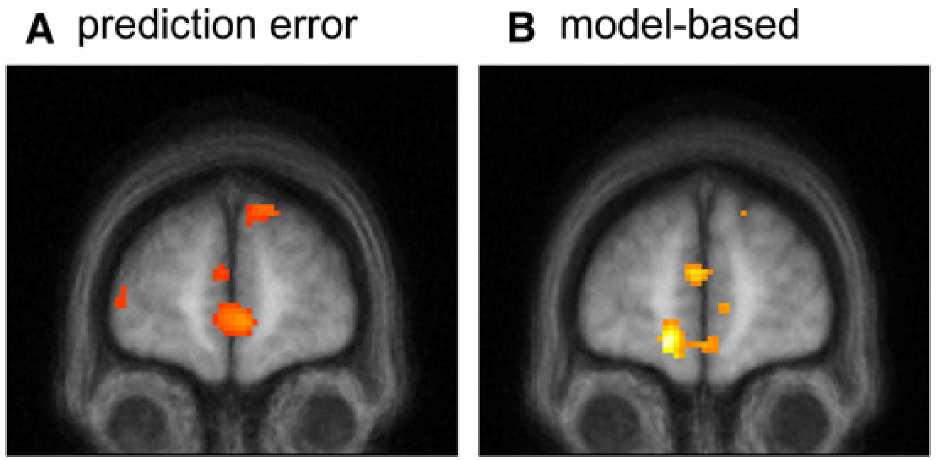
(Daw et al. 2011)





- choice after rare transition

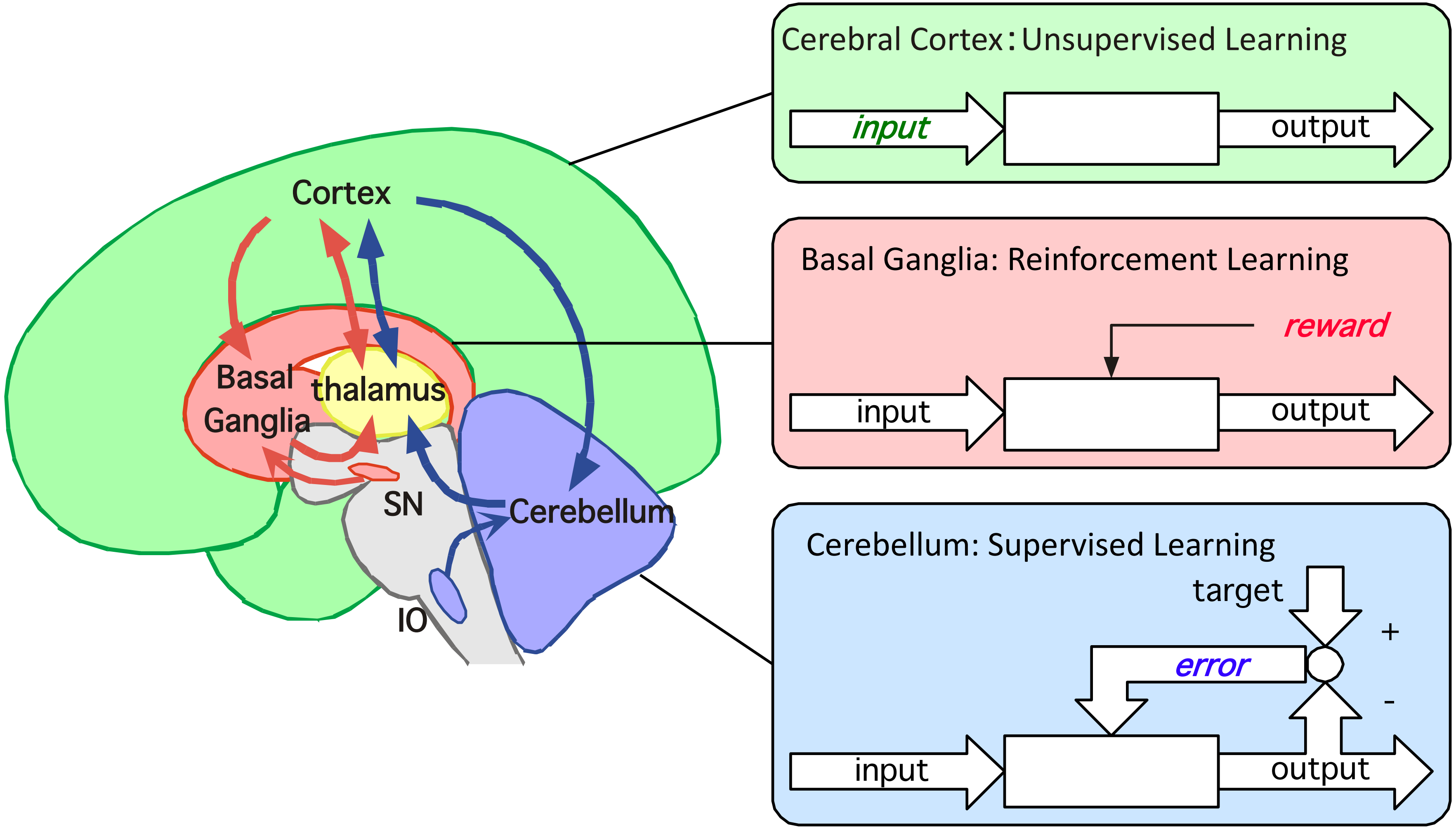




$$Q_{net}(s_A, a_j) = wQ_{MB}(s_A, a_j) + (1-w)Q_{TD}(s_A, a_j)$$

Specialization by Learning Algorithms

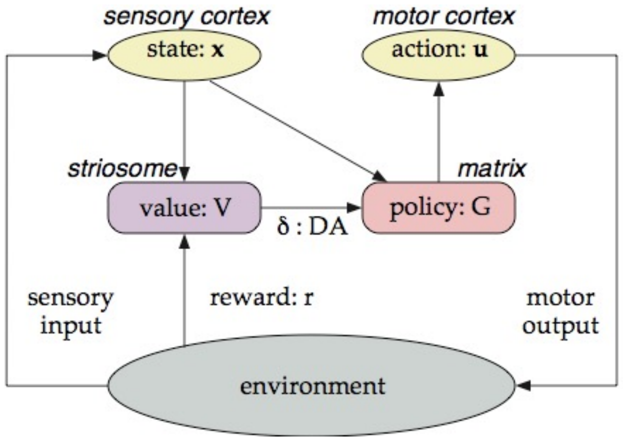
(Doya, 1999)

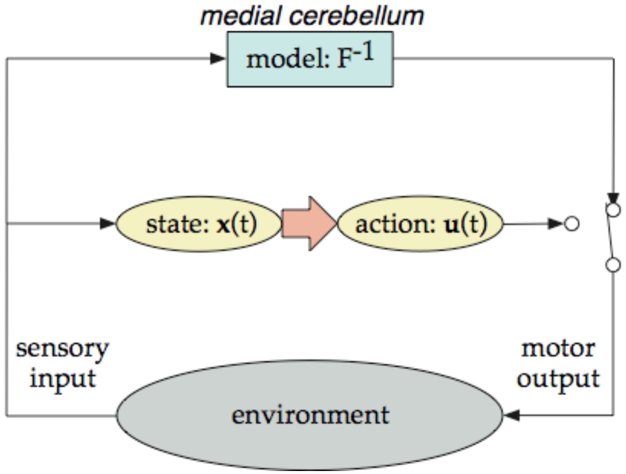


Multiple Action Selection Schemes

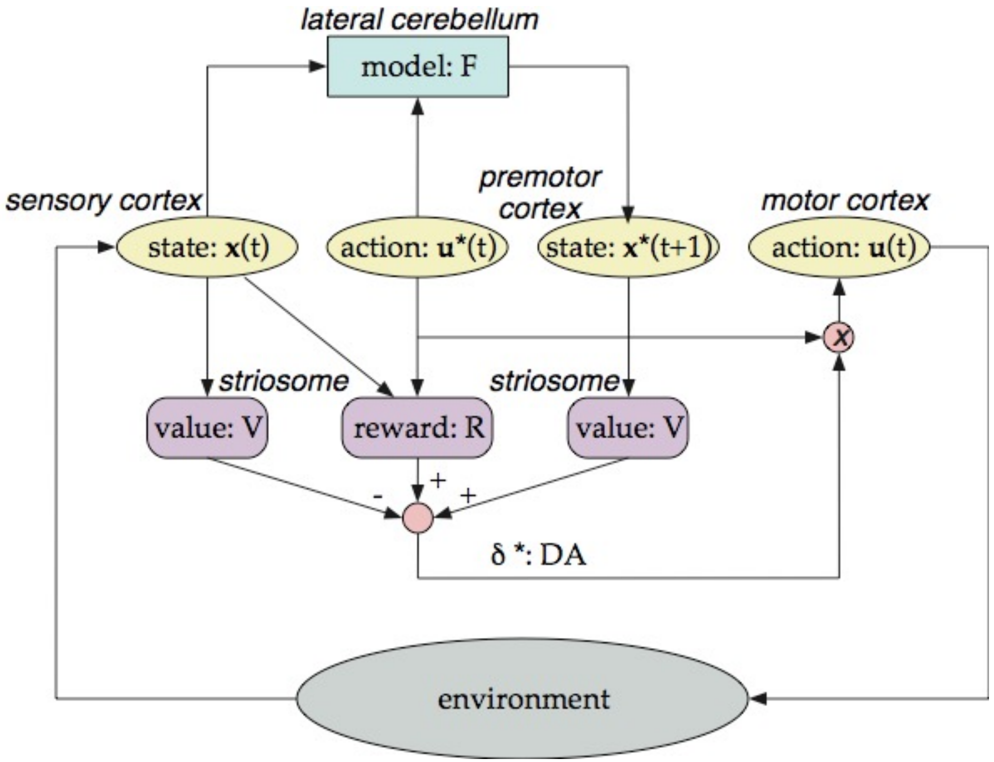
(Doya, 1999)

Model-free





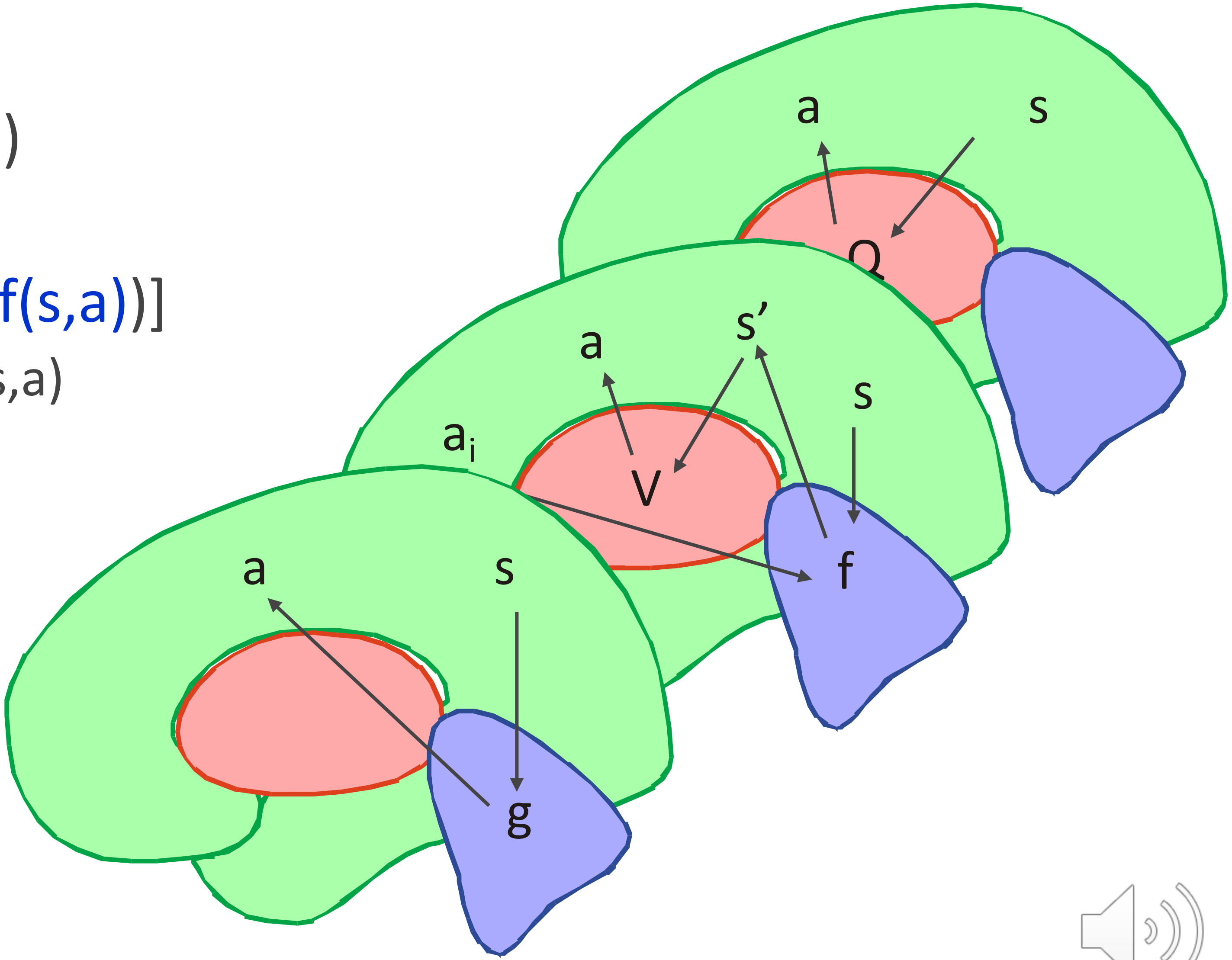
Model-based



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
  - $a = g(s)$



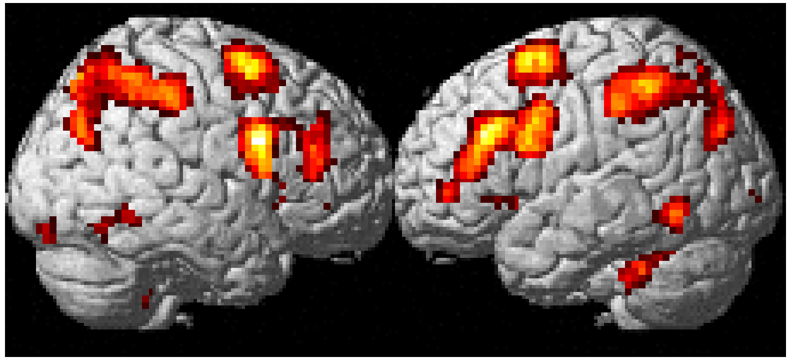
SCIENTIFIC REPORTS

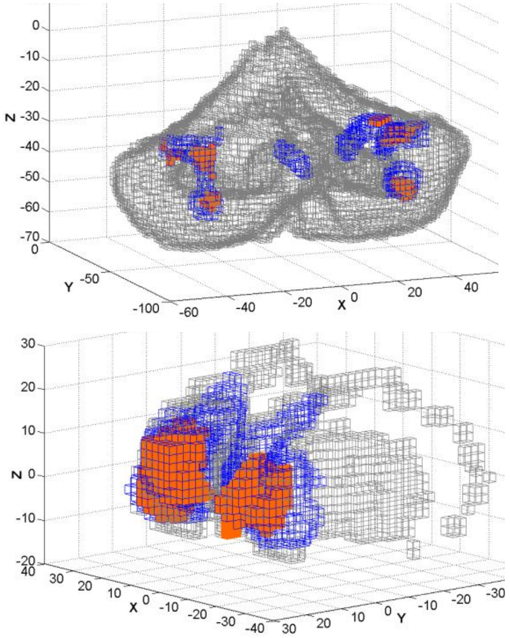
OPEN

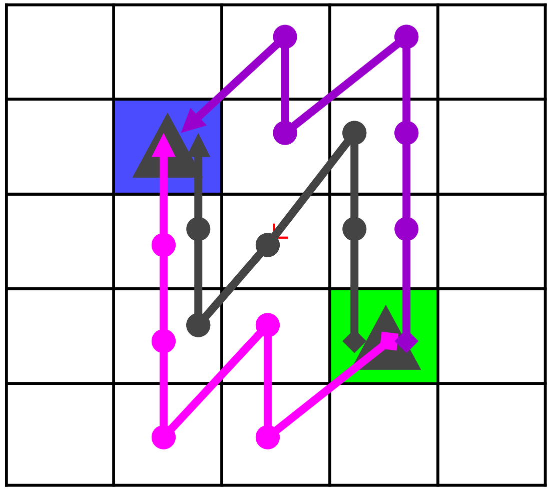
Model-based action planning involves cortico-cerebellar and basal ganglia networks

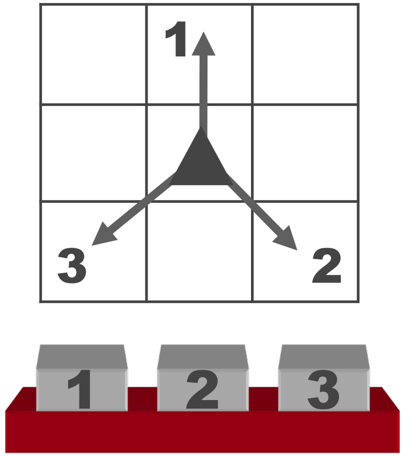
Received: 16 February 2016
Accepted: 19 July 2016

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]





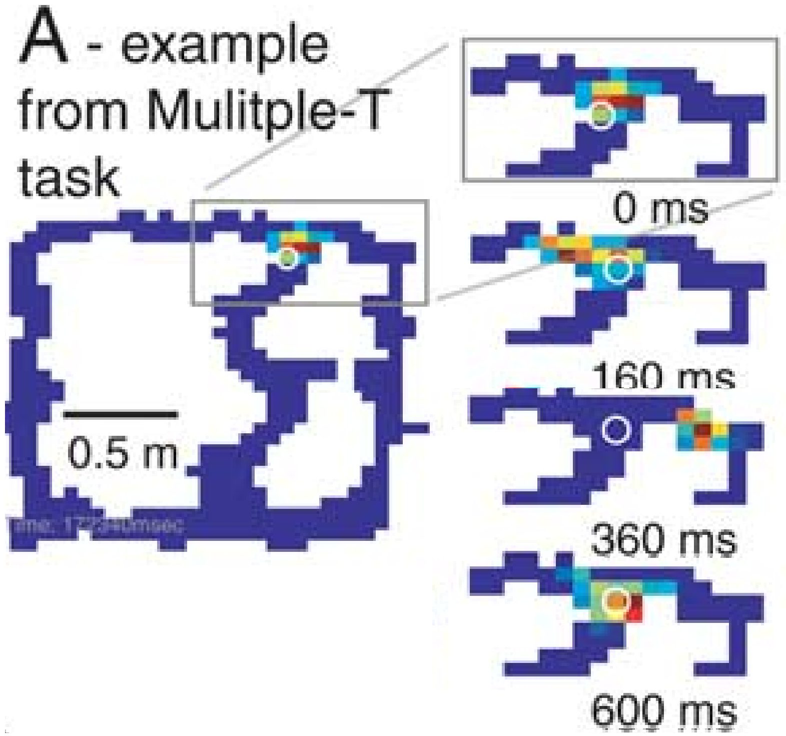




Neuronal Correlates of Mental Simulation

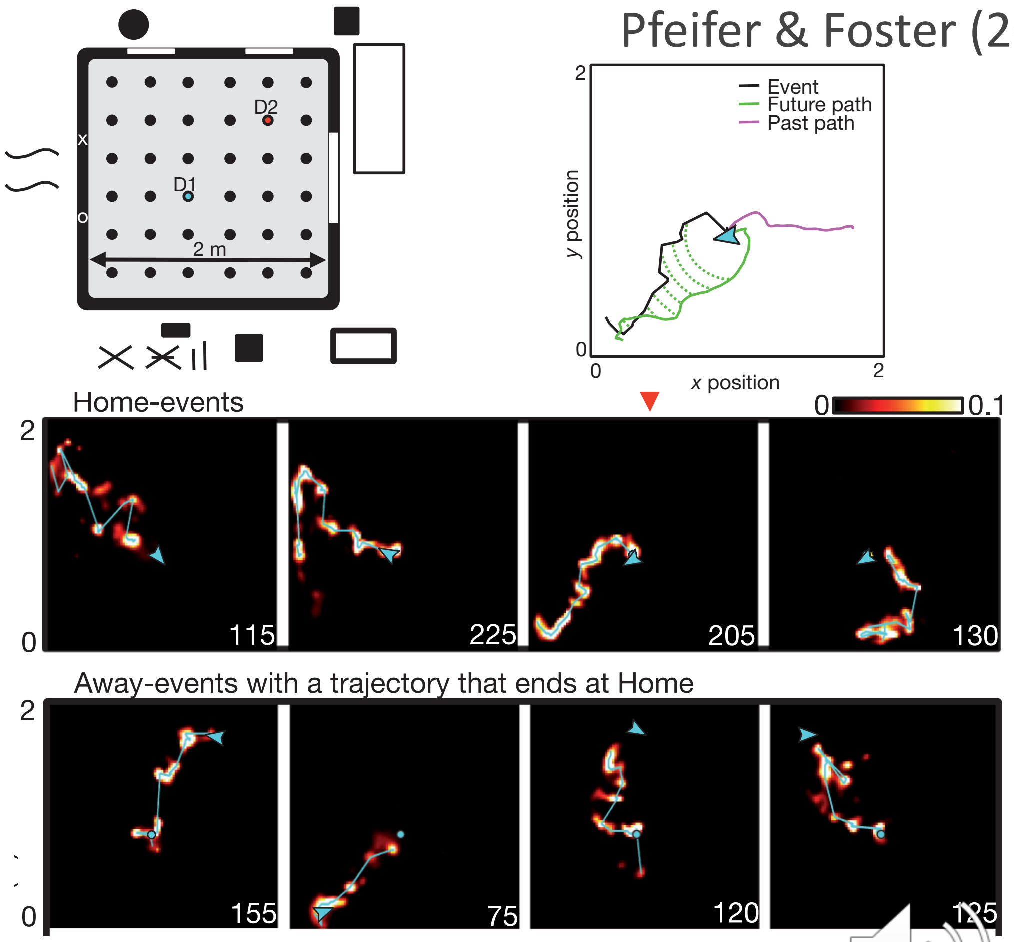
- T-maze

Johnson & Redish (2007)



- Home-Away task

Pfeifer & Foster (2013)

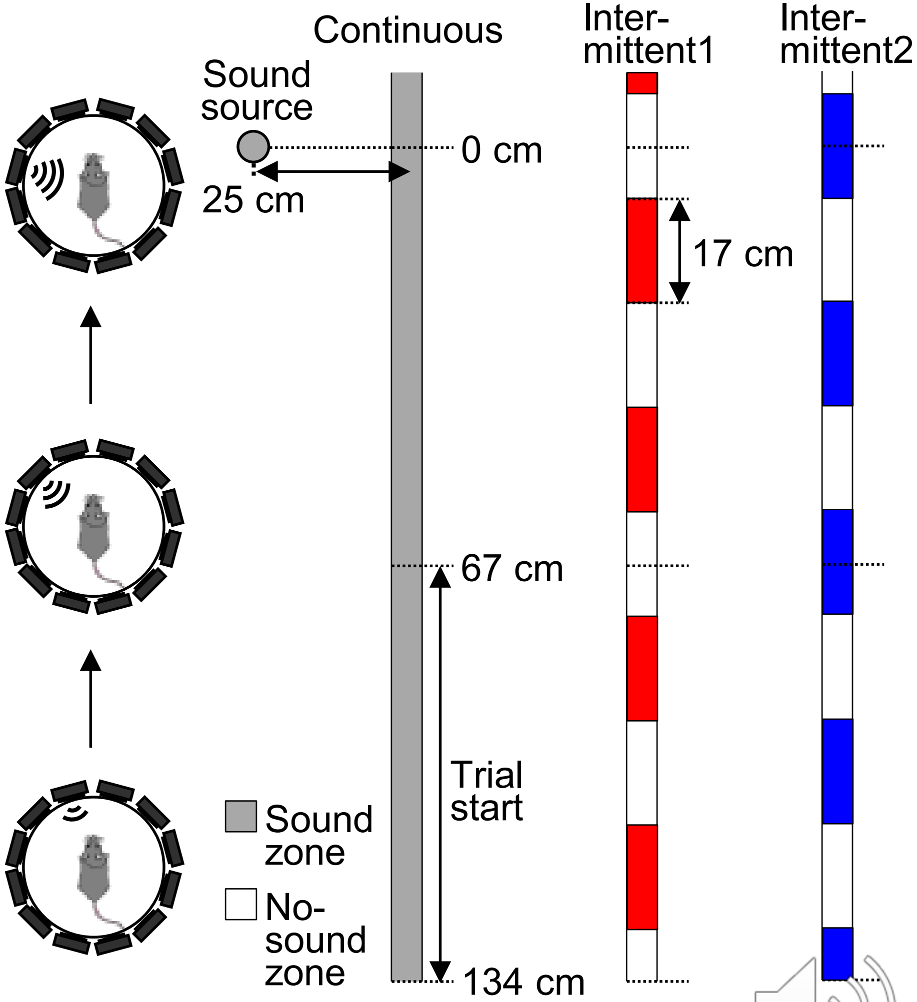


nature neuroscience

Neural substrate of dynamic Bayesian inference in the cerebral cortex

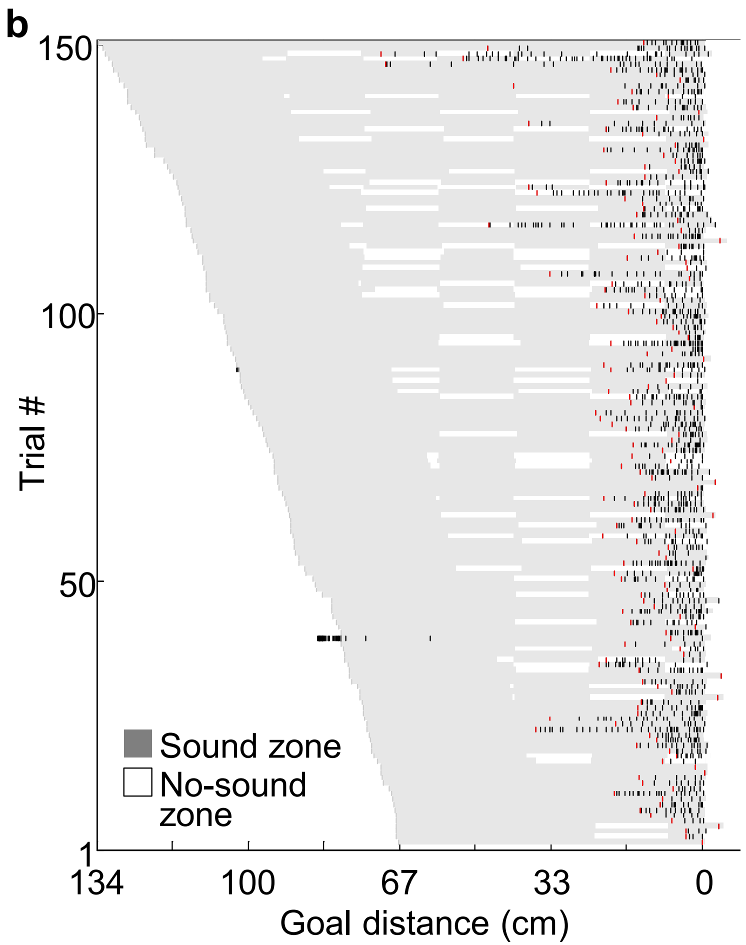
Akihiro Funamizu[1,2], Bernd Kuhn[2] & Kenji Doya[1]

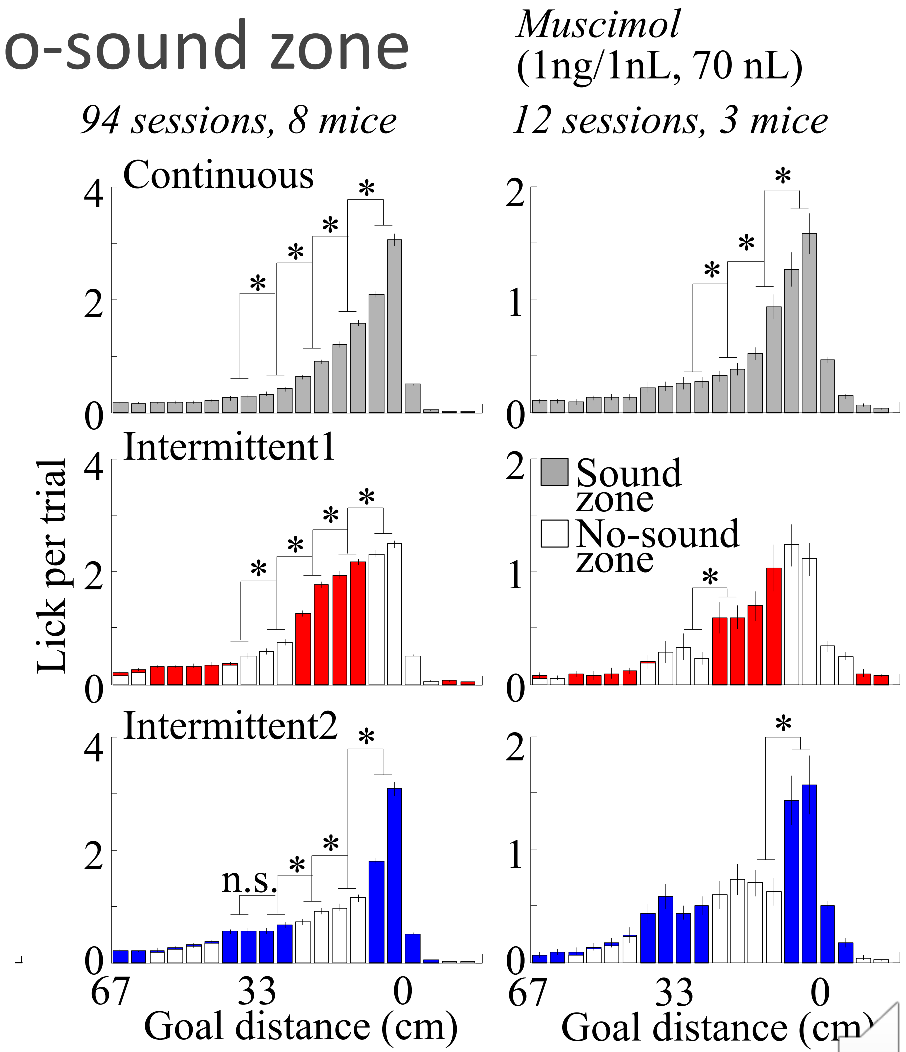
- Auditory virtual environment



Anticipatory Licking

- Mice estimated goal distance in no-sound zone

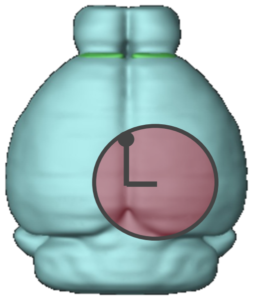


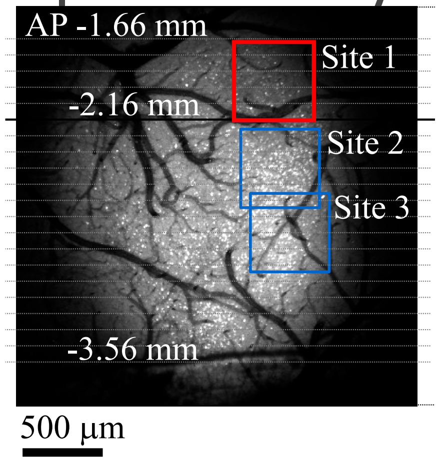


- impaired by muscimol injection in PPC

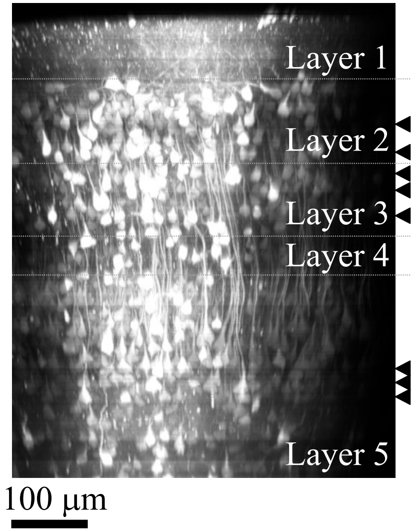
Two-Photon Neural Imaging

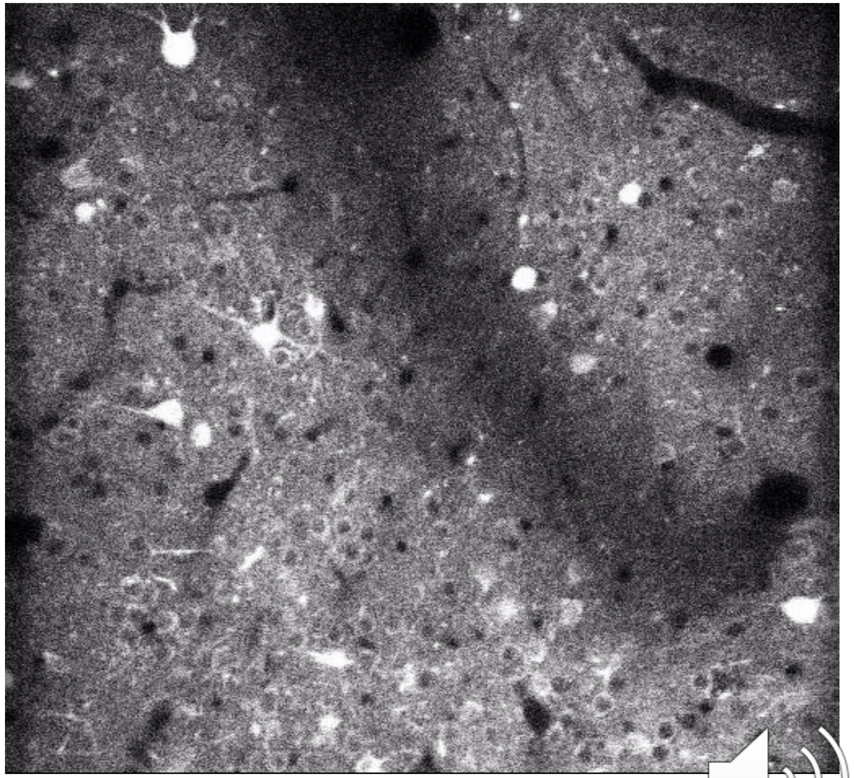
- GCaMP6f expression by AAV



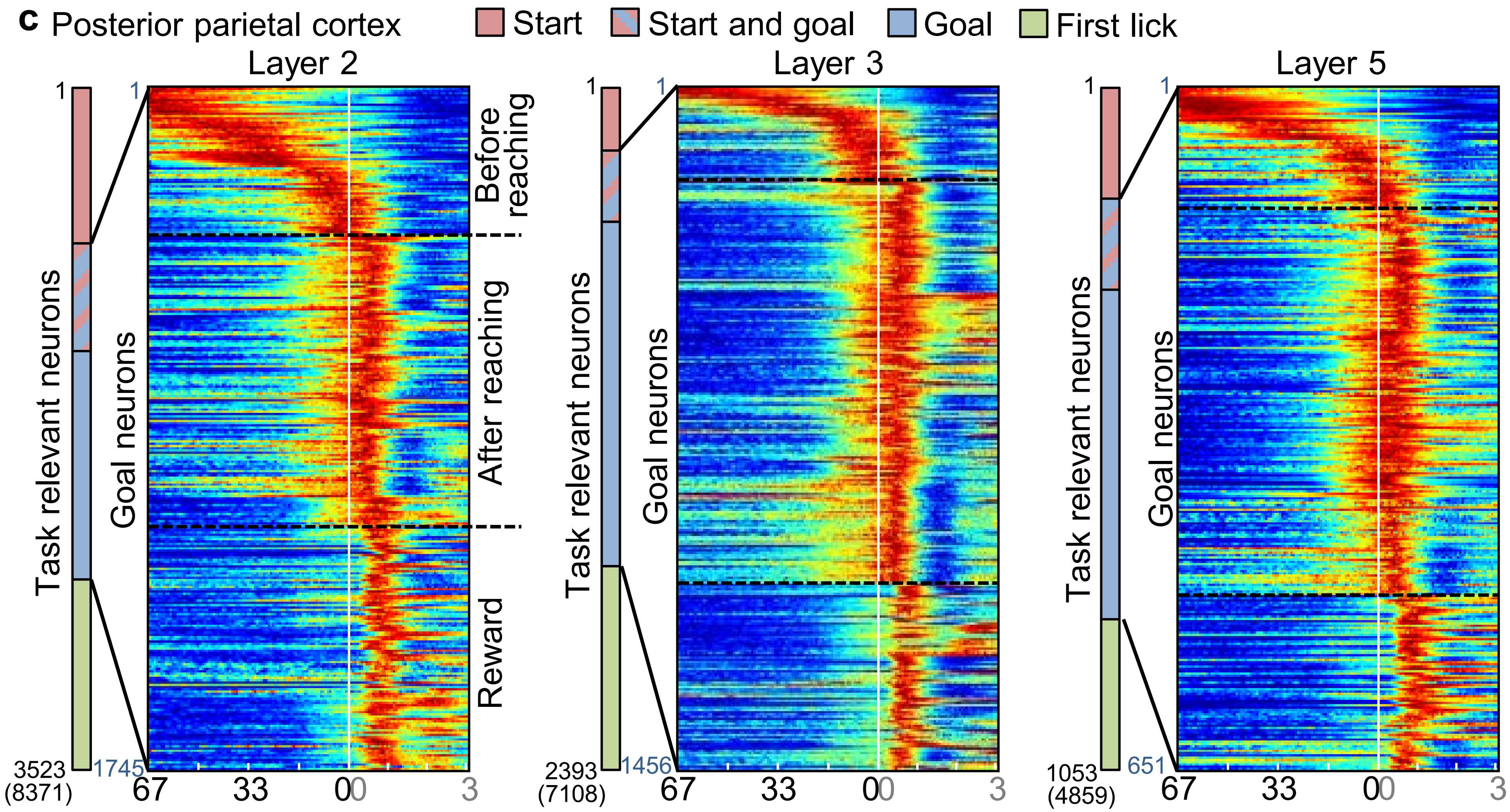


- posterior parietal cortex (PPC)
- auditory-visual cortex (area PM)





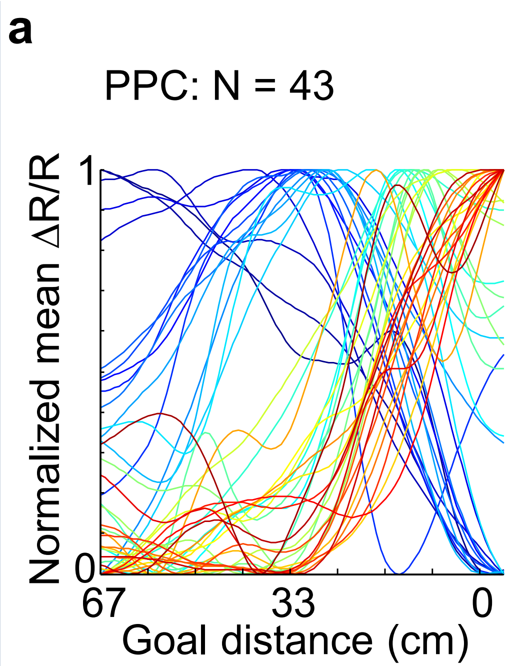
Overall Activities

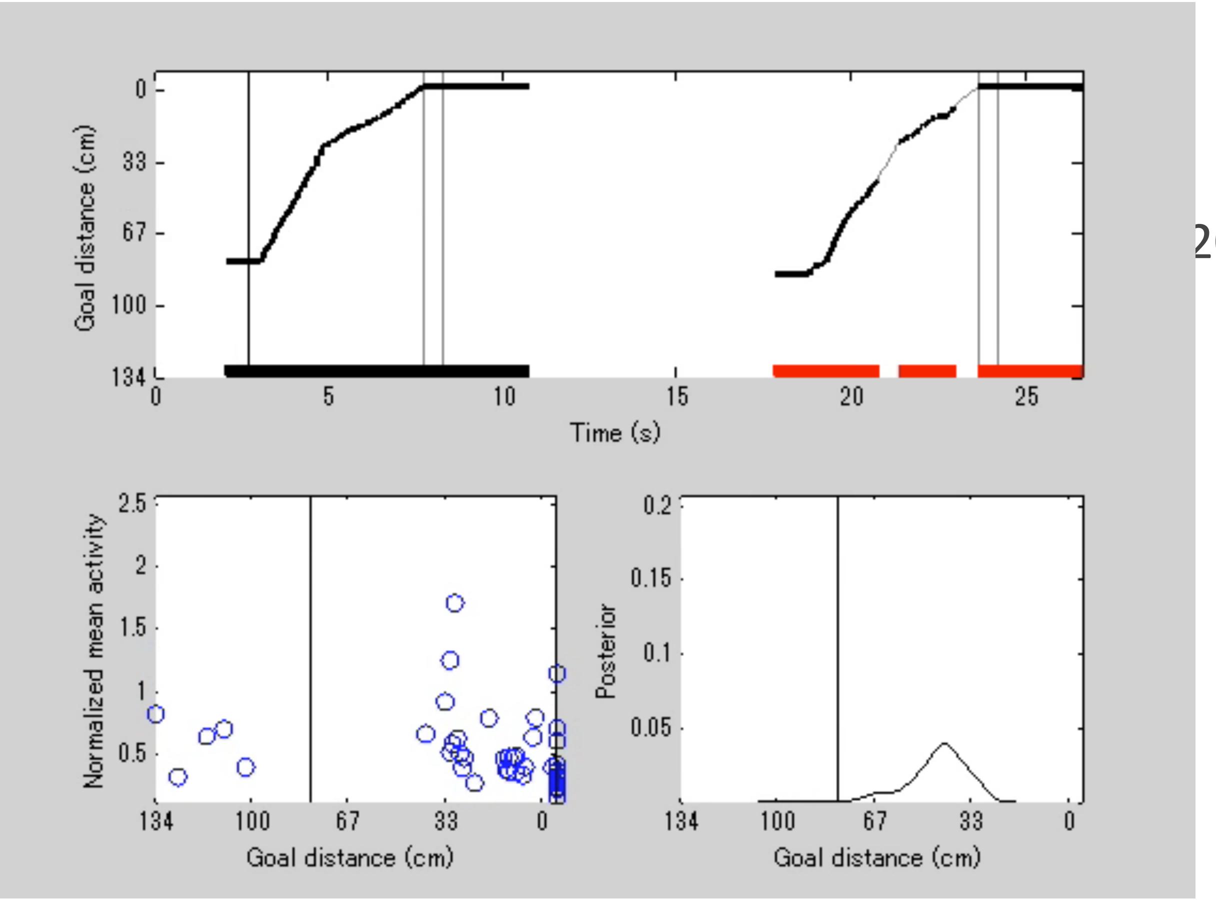


Decoding the Goal Distance

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

2006)

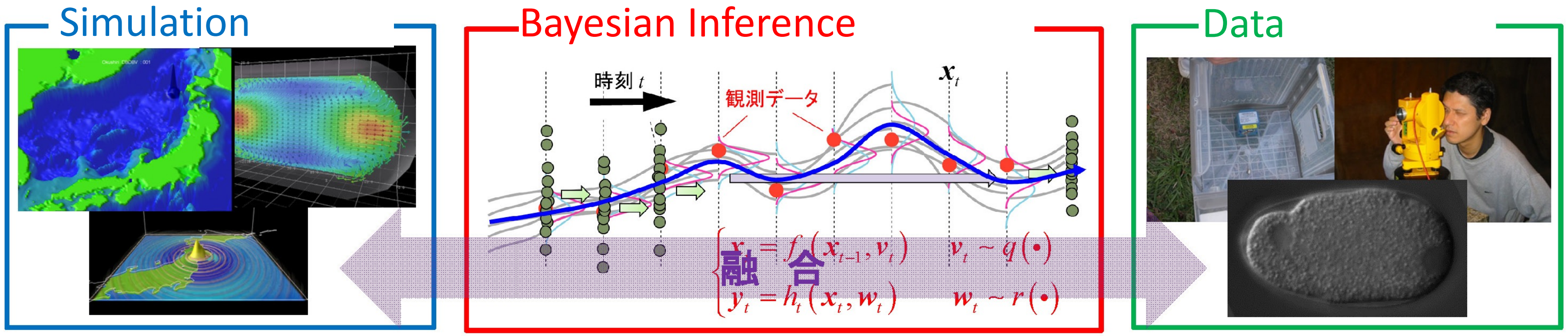




Goal distance (cm)

  - goal distance updated under sound omission

Consciousness as Data Assimilation?



(Komaki Lab. U Tokyo)

- Generative model
  - dynamics $\mathbf{x}_t = f(\mathbf{x}_{t-1}) + \varepsilon_s$ atmosphere, ocean,...
  - observation $\mathbf{y}_t = g(\mathbf{x}_t) + \varepsilon_o$ temperature, wind,...
- State estimation by real-time simulation
  - utilizing sparse, multi-modal data
- Prediction of future states
- Postdiction of past history

Characteristics akin to consciousness

- reason for the emergence of conscious phenomenology?

Kalman's Duality

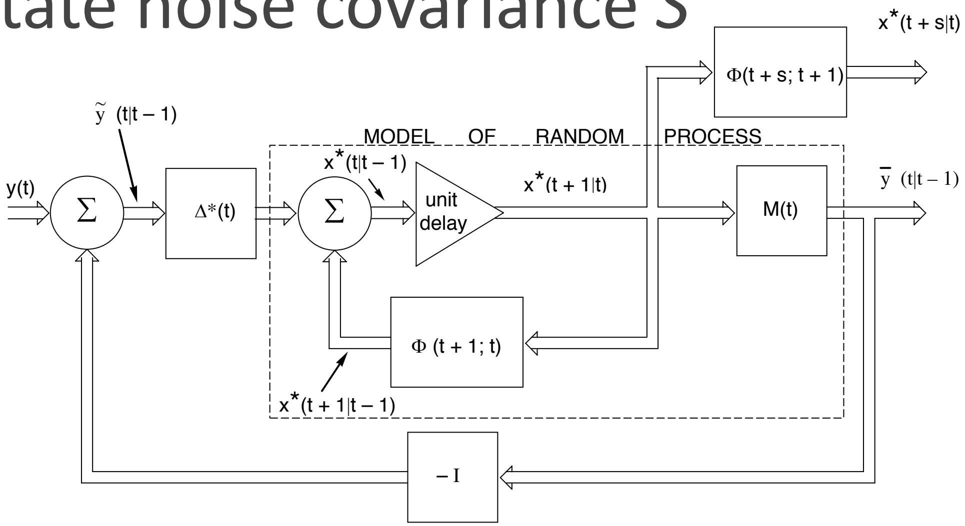
- Optimal filter
- Optimal control

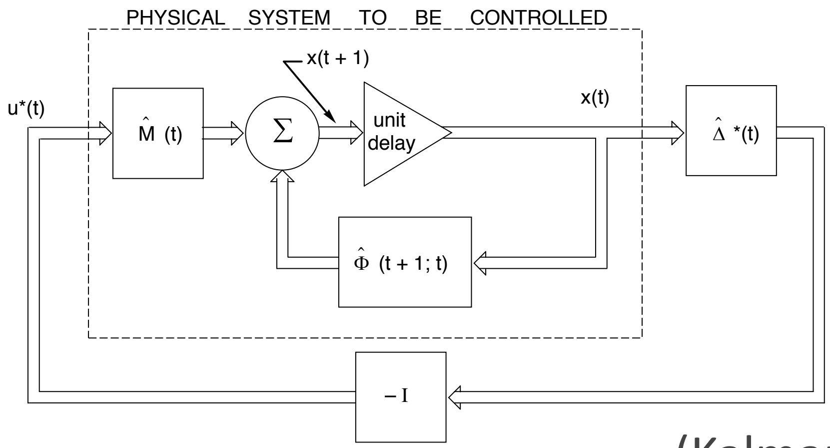
$$\Sigma_{k+1} = S + A\Sigma_k A^\top - A\Sigma_k H^\top \left(P + H\Sigma_k H^\top\right)^{-1} H\Sigma_k A^\top$$

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

- state covariance Σ
- observation gain H
- observation noise covariance P
- state noise covariance S

- quadratic state value matrix V
- action gain B
- action cost matrix R
- state cost matrix Q





(Kalman, 1960)

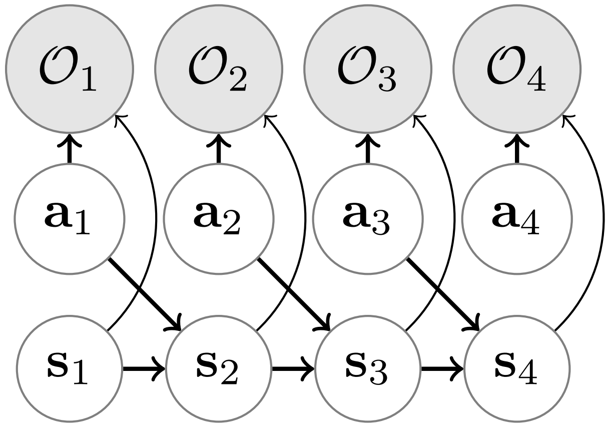
Reinforcement Learning and Control as Probabilistic Inference: Tutorial and Review

(Levine 2018)

Sergey Levine
UC Berkeley

Optimality variable

$$p(\mathcal{O}_t = 1|\mathbf{s}_t, \mathbf{a}_t) = \exp(r(\mathbf{s}_t, \mathbf{a}_t)).$$



Posterior of trajectory for *O*= 1

$$p(\tau|\mathbf{o}_{1:T}) \propto \mathbb{1}[p(\tau) \neq 0] \exp\left(\sum_{t=1}^{T} r(\mathbf{s}_t, \mathbf{a}_t)\right)$$

- optimal policy

$$p(\mathbf{a}_t|\mathbf{s}_t, \mathcal{O}_{t:T} = 1)$$

Policy search as inference

- backward message

$$\beta_t(\mathbf{s}_t, \mathbf{a}_t) = p(\mathcal{O}_{t:T}|\mathbf{s}_t, \mathbf{a}_t)$$

$$\beta_t(\mathbf{s}_t) = p(\mathcal{O}_{t:T}|\mathbf{s}_t) = \int_{\mathcal{A}} p(\mathcal{O}_{t:T}|\mathbf{s}_t, \mathbf{a}_t)p(\mathbf{a}_t|\mathbf{s}_t)d\mathbf{a}_t = \int_{\mathcal{A}} \beta_t(\mathbf{s}_t, \mathbf{a}_t)p(\mathbf{a}_t|\mathbf{s}_t)d\mathbf{a}_t$$

$$\beta_t(\mathbf{s}_t, \mathbf{a}_t) = p(\mathcal{O}_{t:T}|\mathbf{s}_t, \mathbf{a}_t) = \int_{\mathcal{S}} \beta_{t+1}(\mathbf{s}_{t+1})p(\mathbf{s}_{t+1}|\mathbf{s}_t, \mathbf{a}_t)p(\mathcal{O}_t|\mathbf{s}_t, \mathbf{a}_t)d\mathbf{s}_{t+1}$$

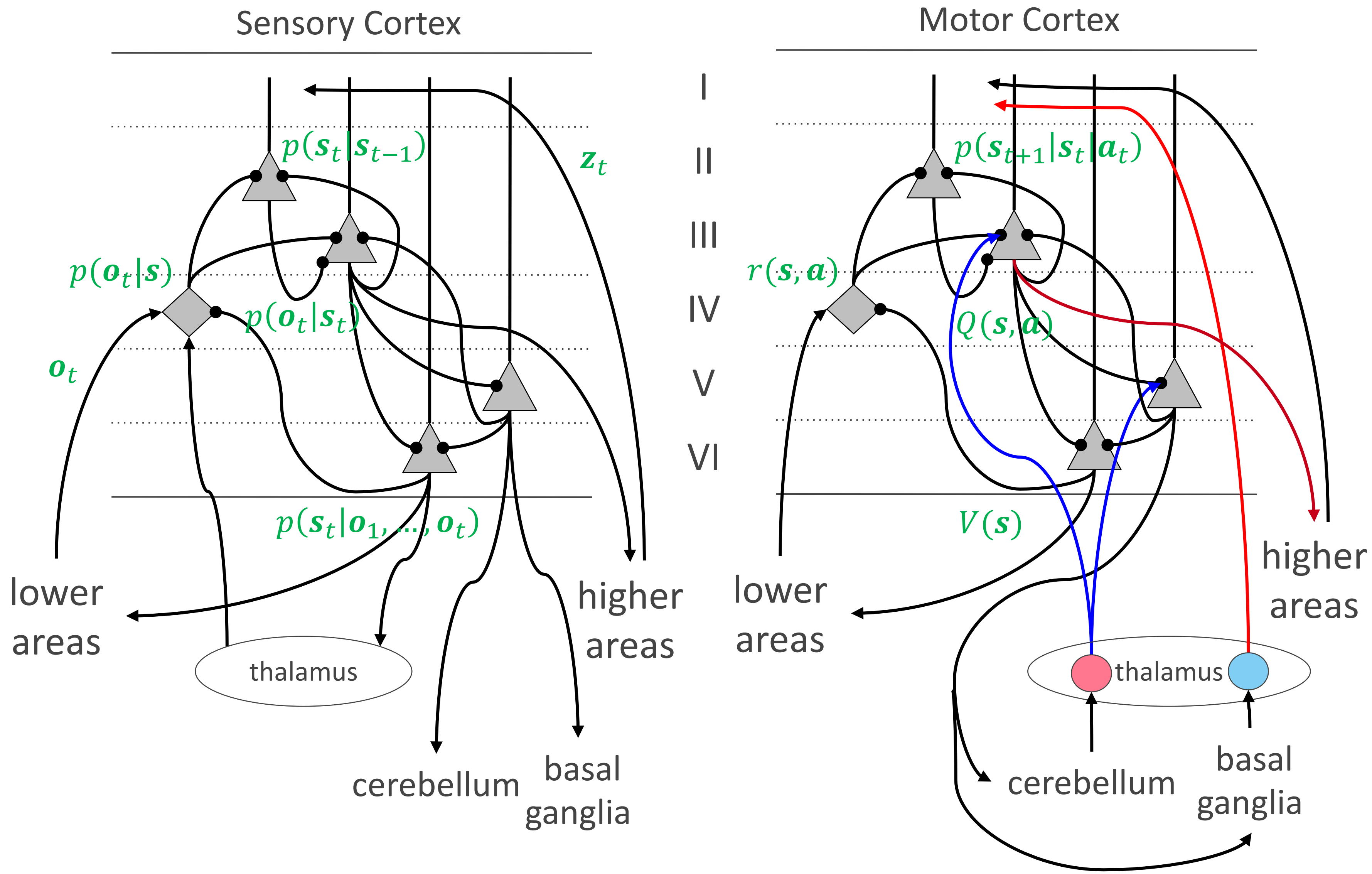
- optimal poicy

$$p(\mathbf{a}_t|\mathbf{s}_t, \mathcal{O}_{t:T}) = \frac{\beta_t(\mathbf{s}_t, \mathbf{a}_t)}{\beta_t(\mathbf{s}_t)}$$

- value functions

$$Q(\mathbf{s}_t, \mathbf{a}_t) = \log \beta_t(\mathbf{s}_t, \mathbf{a}_t)$$
$$V(\mathbf{s}_t) = \log \beta_t(\mathbf{s}_t).$$

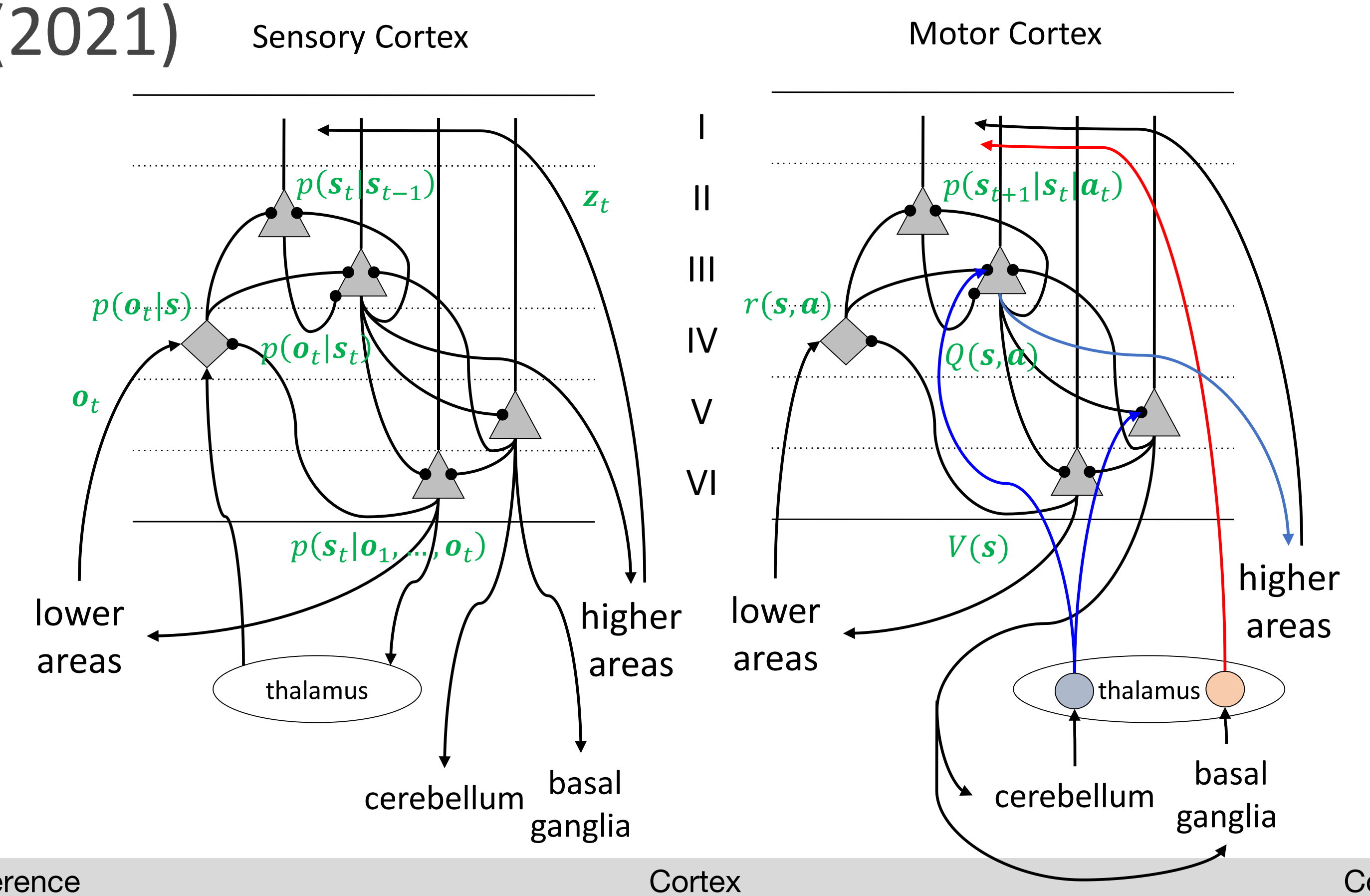
Canonical Cortical Circuits



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

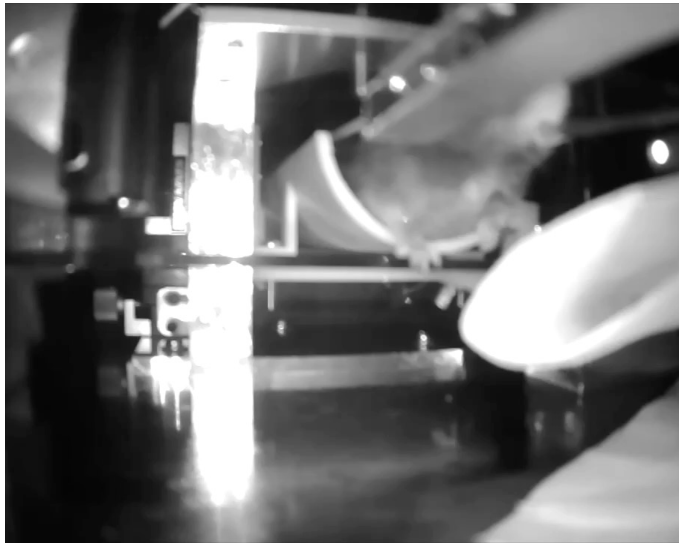
Kenji Doya (2021)

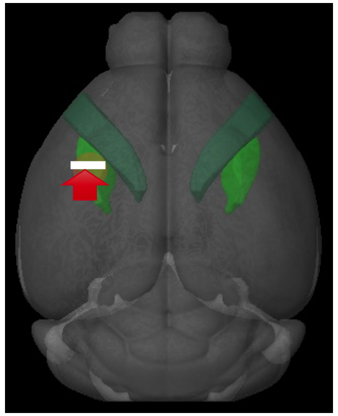
Current Opinion in Behavioral Sciences

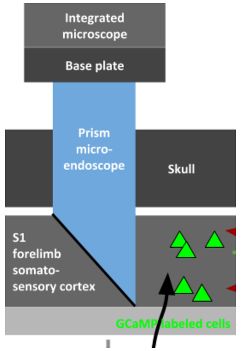


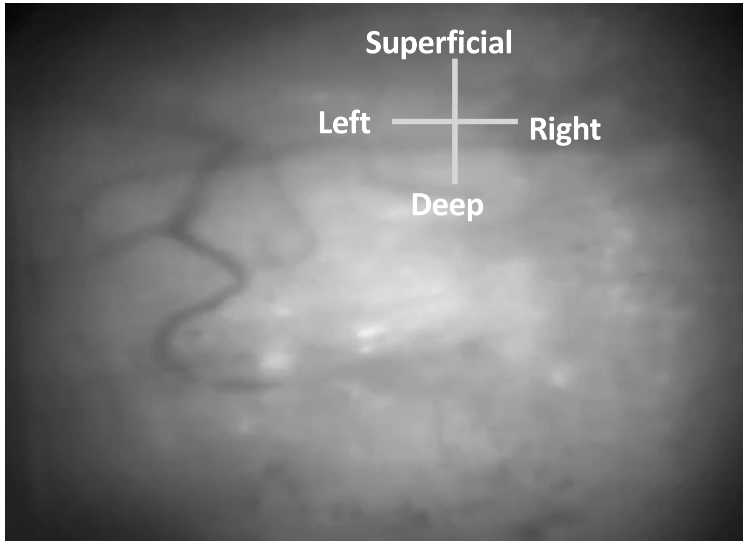
Inference	Cortex	Control
Top-down signal $\boldsymbol{z}_t$	L1 input	Top-down activation signal
Bottom-up signal $p(\boldsymbol{o}_t\|\boldsymbol{s}_t)$	L2/3 output	Action value $Q(\boldsymbol{s},\boldsymbol{a})$
Predictive model $p(\boldsymbol{s}_t\|\boldsymbol{s}_{t-1})$	L2/3 connection	Predictive model $p(\boldsymbol{s}_{t+1}\|\boldsymbol{s}_t,\boldsymbol{a}_t)$
Bottom-up signal $\boldsymbol{o}_t$	L4 input	Optimality signal $\mathcal{O}_t$
Likelihood $p(\boldsymbol{o}_t\|\boldsymbol{s})$	L4 output	Reward function $r(\boldsymbol{s},\boldsymbol{a})$
Posterior $p(\boldsymbol{s}_t\|\boldsymbol{o}_1,\ldots,\boldsymbol{o}_t)$	L5 output	State value $V(\boldsymbol{s})$
Top-down signal $\boldsymbol{s}_t$	L6 output	Action $p(\boldsymbol{a}_t\|\boldsymbol{s}_t)$

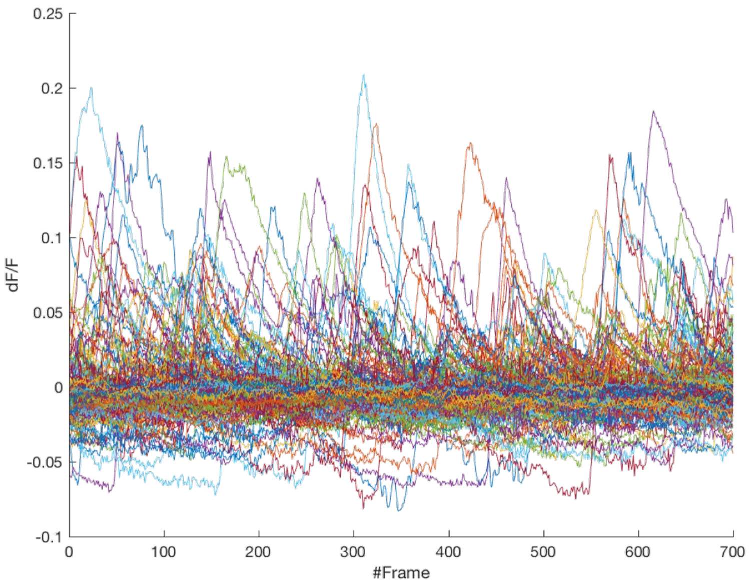
Prism Lens Imaging during Lever Pull Task

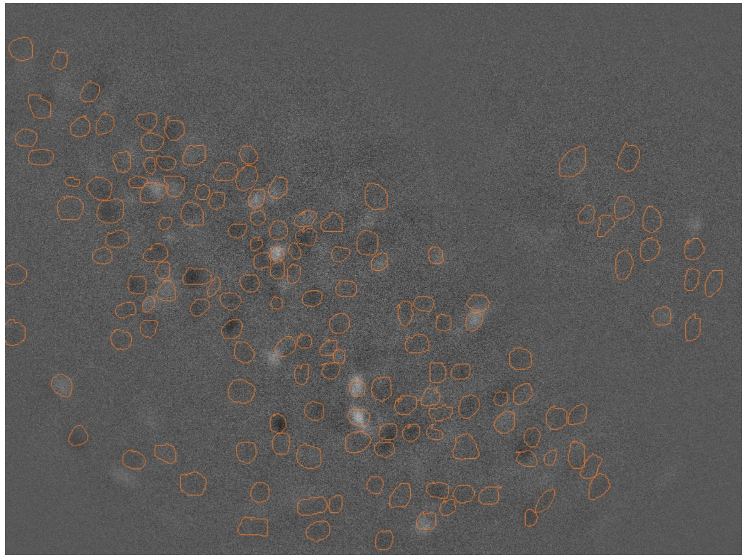












Reinforcement Learning

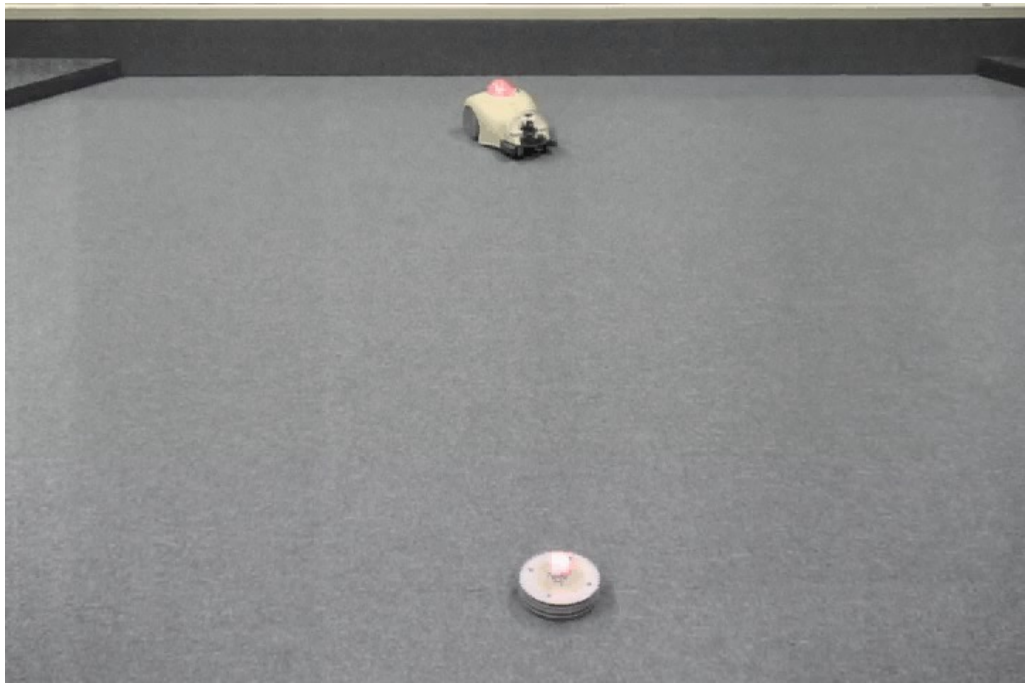
- Predict reward: *value function*
  - $V(s) = E[\, r(t) + \gamma r(t+1) + \gamma^2 r(t+2) \ldots \mid s(t)=s]$
  - $Q(s,a) = E[\, r(t) + \gamma r(t+1) + \gamma^2 r(t+2) \ldots \mid s(t)=s,\ a(t)=a]$
- Select action
  - *greedy*: $a = \text{argmax}\ Q(s,a)$
  - *Boltzmann*: $P(a|s) \propto \exp[\,\beta\,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

How to implement these steps?

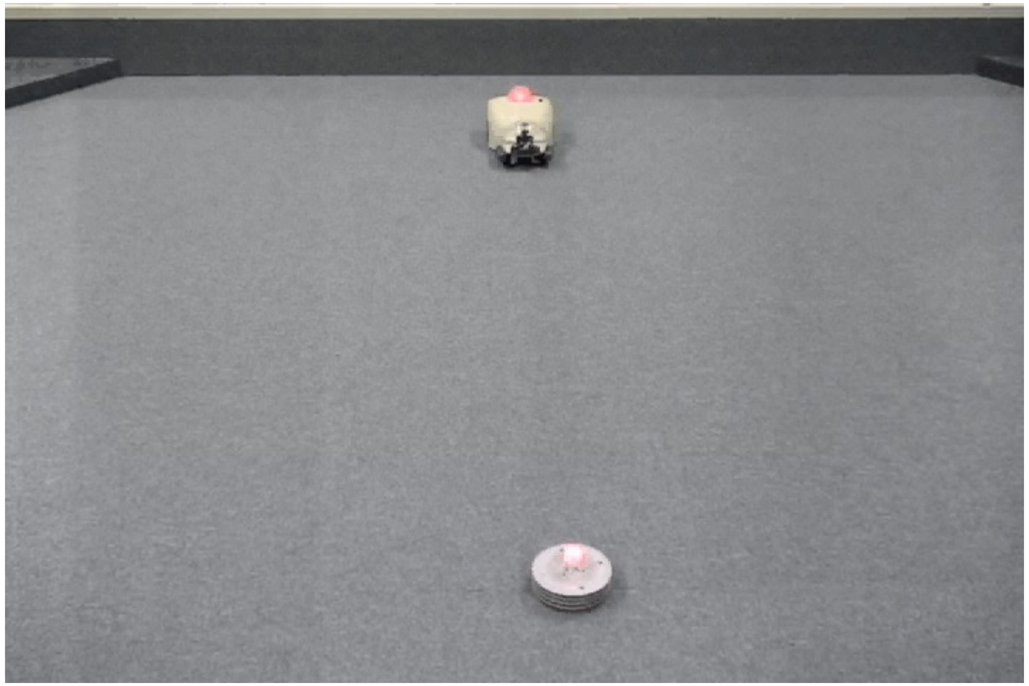
How to tune these parameters?

Temporal Discount Factor γ

- Large γ
  - reach for far reward



- Small γ
  - only to near reward



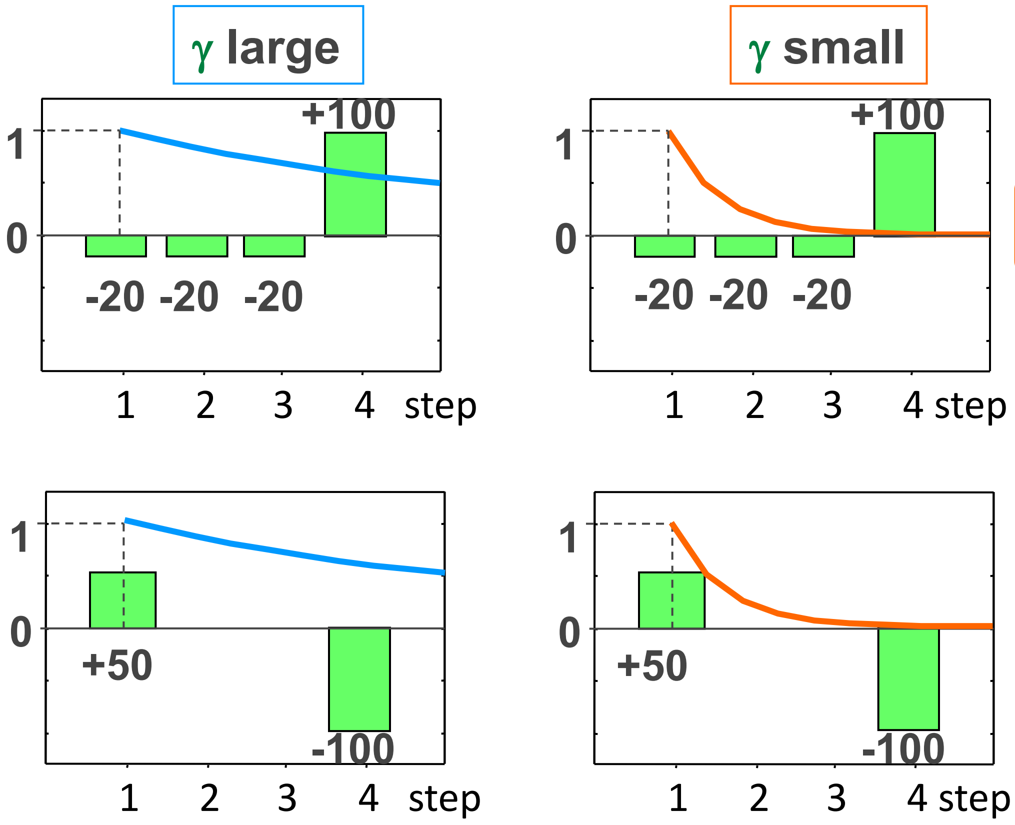
Temporal Discount Factor γ

- $V(t) = E[\, r(t) + \gamma r(t+1) + \gamma^2 r(t+2) + \gamma^3 r(t+3) + \ldots]$
  - 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

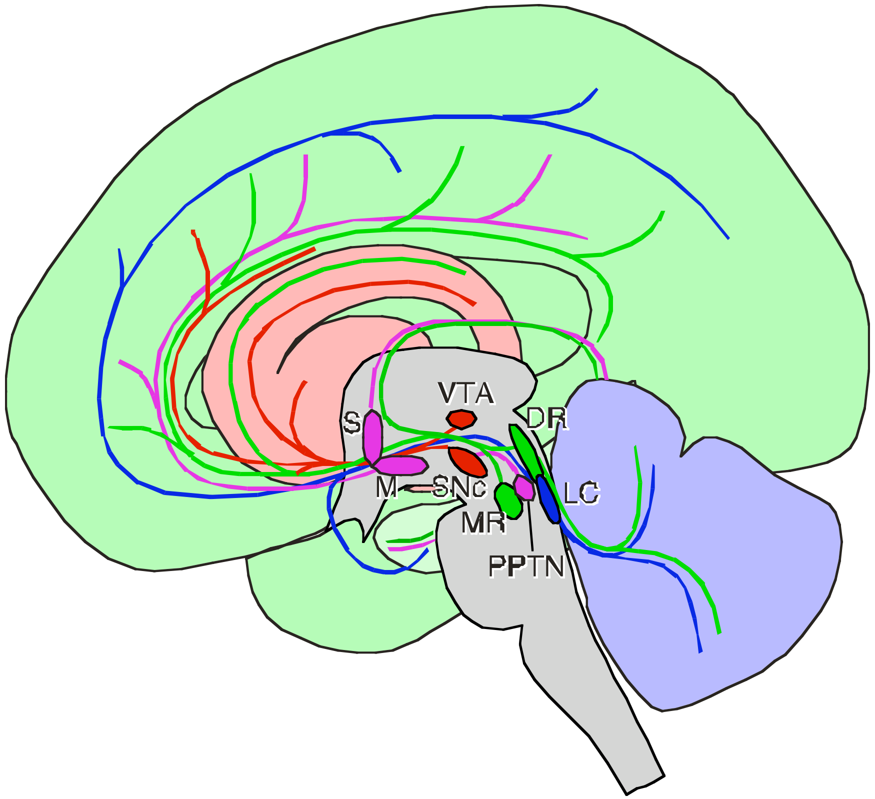
V = 47.3

Serotonin?

Neuromodulators for Metalearning

(Doya, 2002)

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



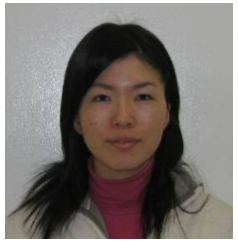
Dopamine: TD error δ

Acetylcholine: learning rate α

Noradrenaline: exploration β

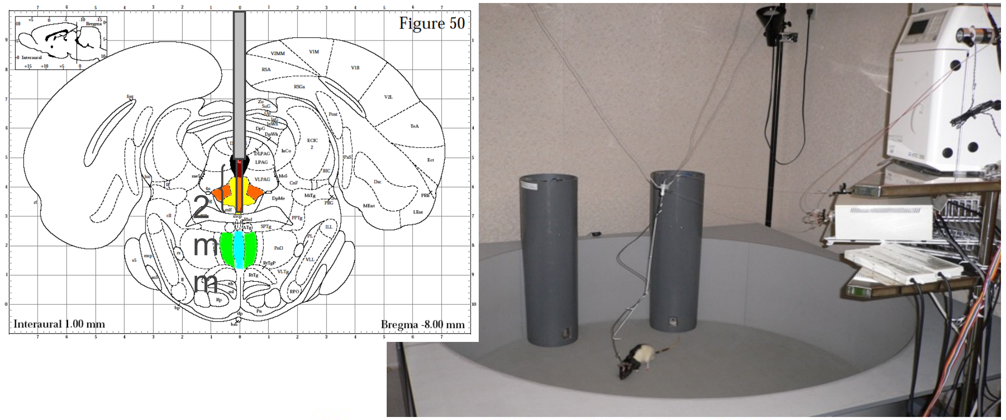
Serotonin: temporal discount γ

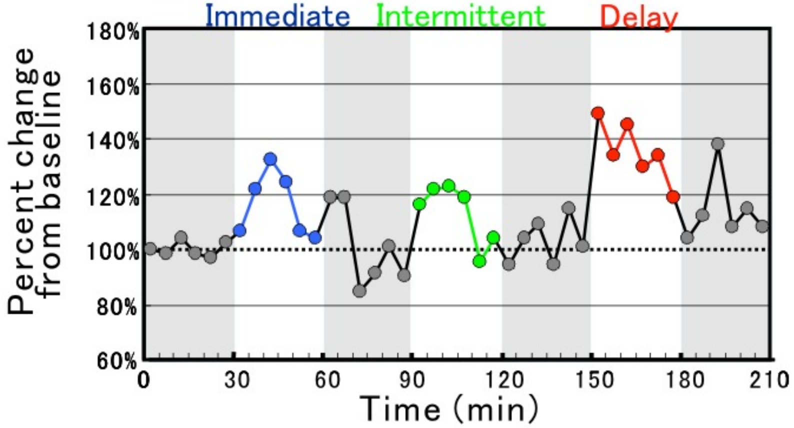
Chemical Measurement/Control



(Kayoko Miyazaki et al., 2011, 2012)

Microdialysis measurement

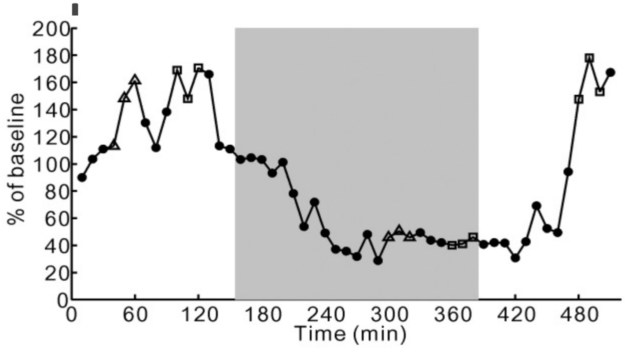


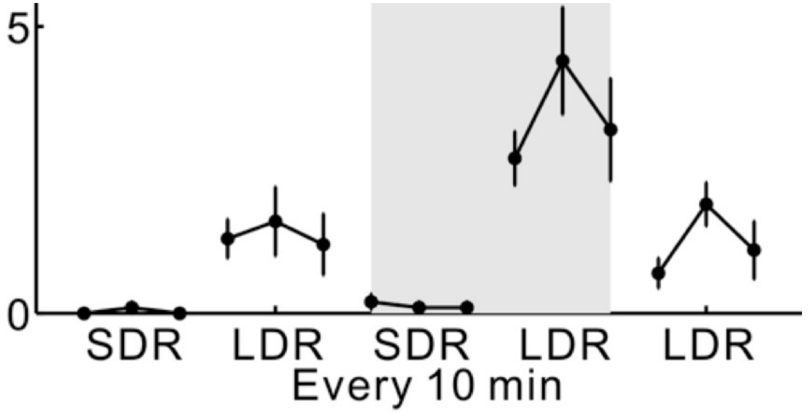


- Serotonin release increased in delayed reward task

Serotonin neuron blockade

- 5HT1A agonist in dorsal r



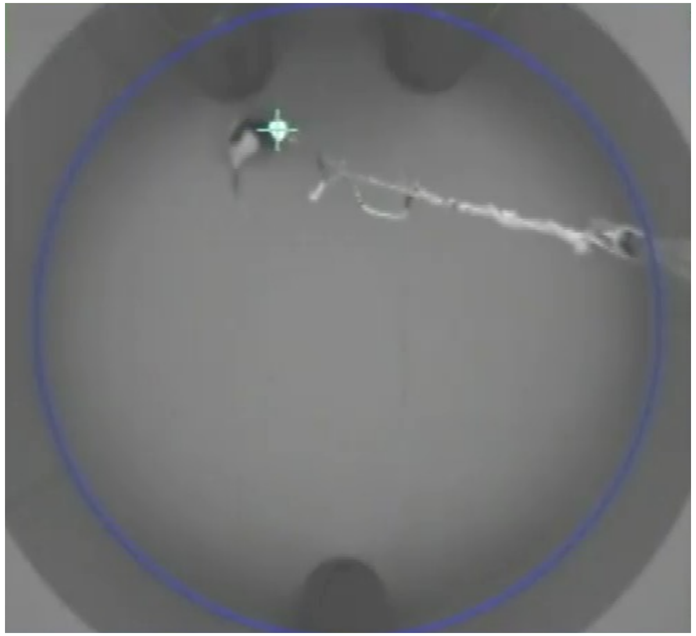


- Waiting error increased in long-delayed reward trials

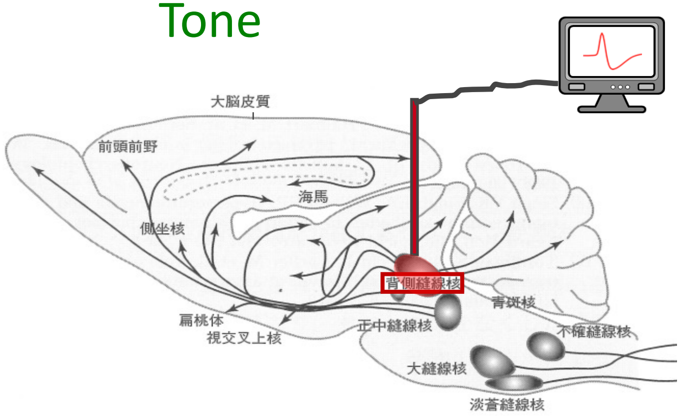
Dorsal Raphe Neuron Recording

(Miyazaki et al. 2011 JNS)

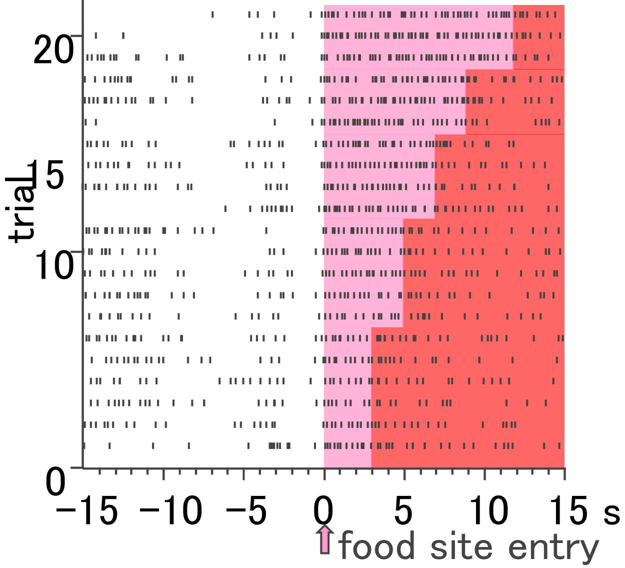
Food Water



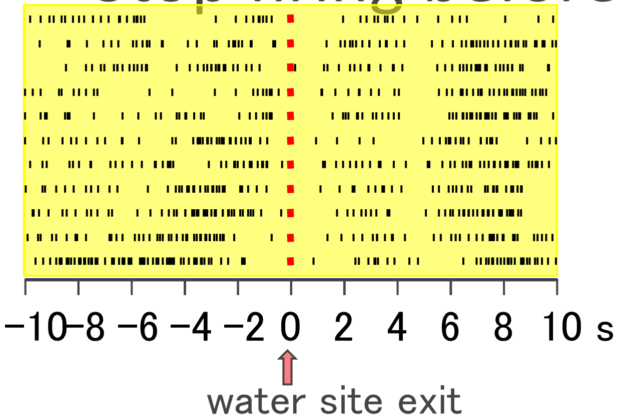
Tone



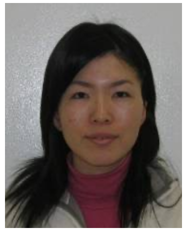
■ Keep firing while waiting



■ Stop firing before giving up



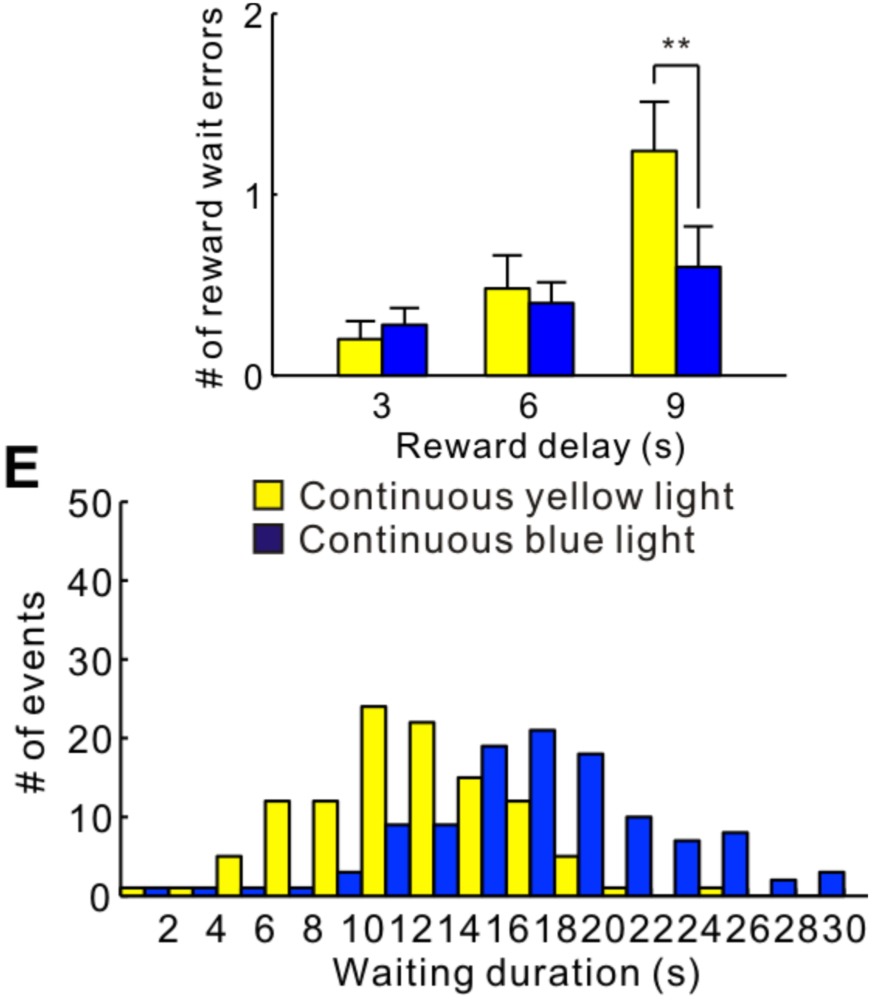
Optogenetic Stimulation of Serotonin Neurons



(Miyazaki et al., 2014, Current Biology)

- Reward Delay Task (3, 6, 9, ∞ sec)

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



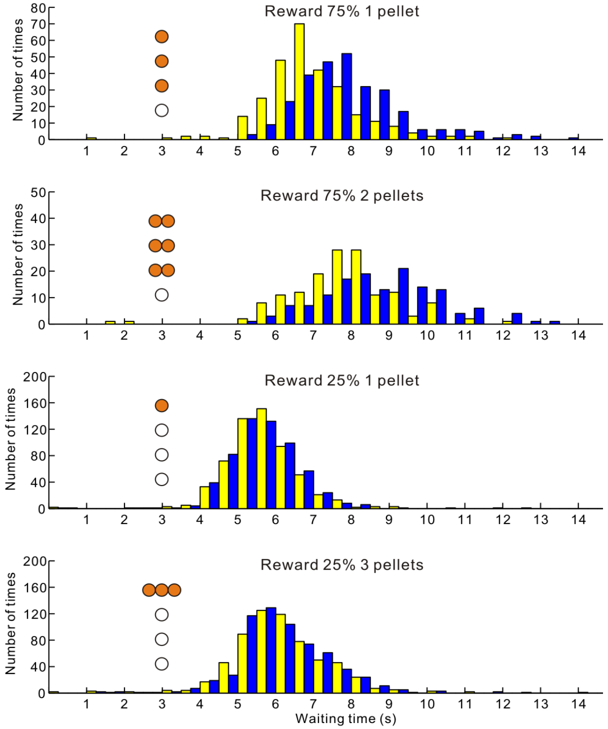
nature COMMUNICATIONS (2018)

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

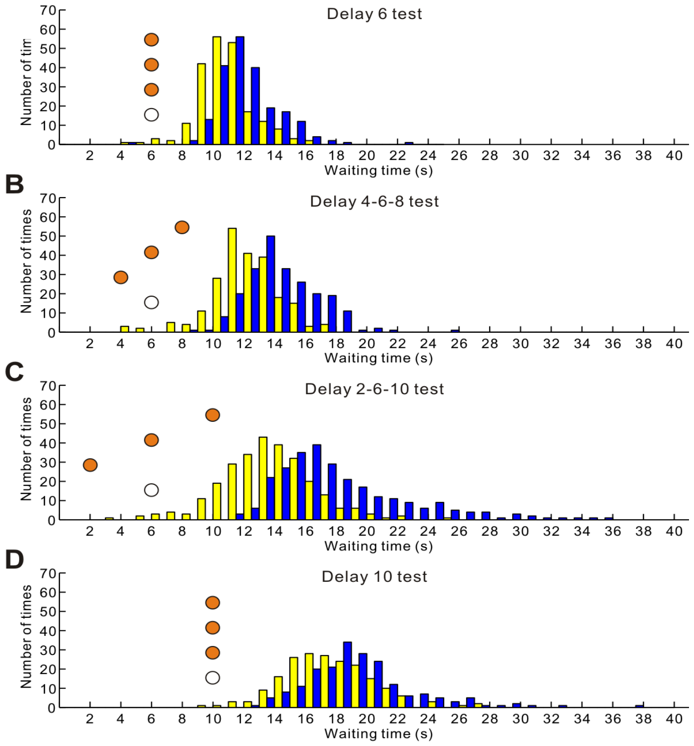
Katsuhiko Miyazaki[1], Kayoko W. Miyazaki[1], Akihiro Yamanaka[2], Tomoki Tokuda[3], Kenji F. Tanaka[4] & Kenji Doya[1]

Serotonin stimulation facilitates waiting when...

- reward delivery is certain

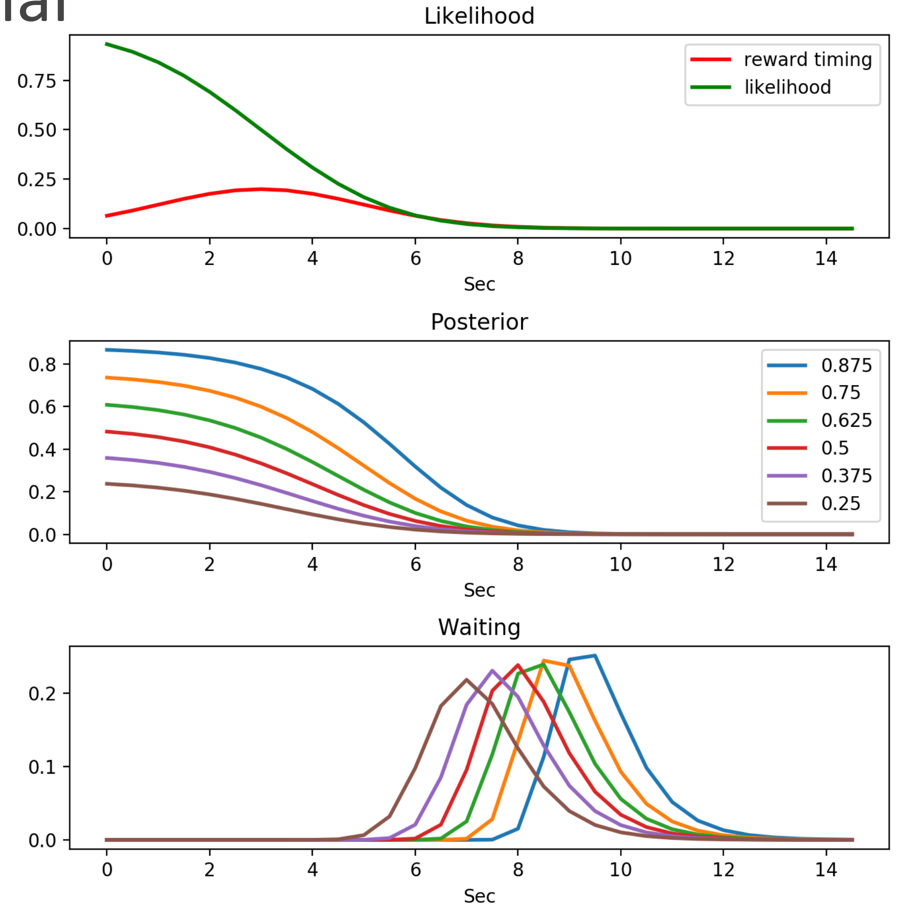


- reward timing is uncertain



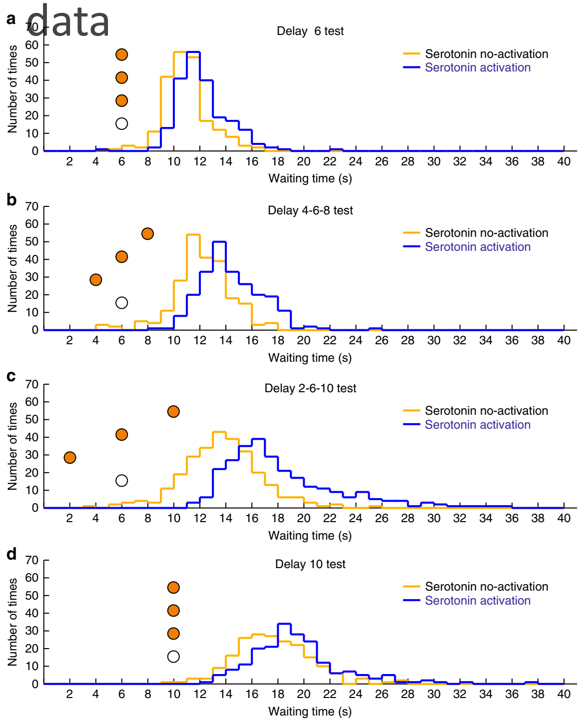
Bayesian Waiting Decision Model

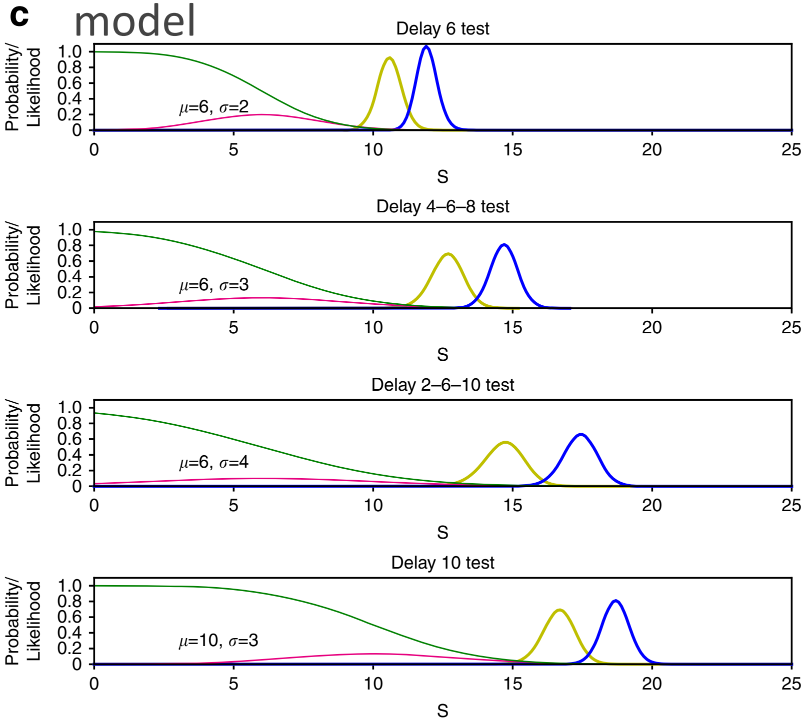
- Mice have internal model of reward timing
  - keep guessing if it is a rewarded trial
- Likelihood of reward drops
  - higher prior sustains posterior
  - timing uncertainty makes long-tailed likelihood
- Serotonin signal reward prior?
  - average reward response (Cohen et al., 2015)



Effect of Timing Uncertainty

- 5-HT stimulation causes longer waiting when reward timing is more uncertain.
- Bayesian model replicates the effect by assuming that 5-HT enhances prior probability of reward.





F. A. Q.

There are so many receptors.
How can serotonin have a single function?

- For the same broadcast message, correct response may be different depending on positions.
  - exchange rate → import/export businesses
- That might be why so many receptors were evolved.

There are multiple origins: DRN (d/v/m/l), MRN, descending, …; do they carry the same message?

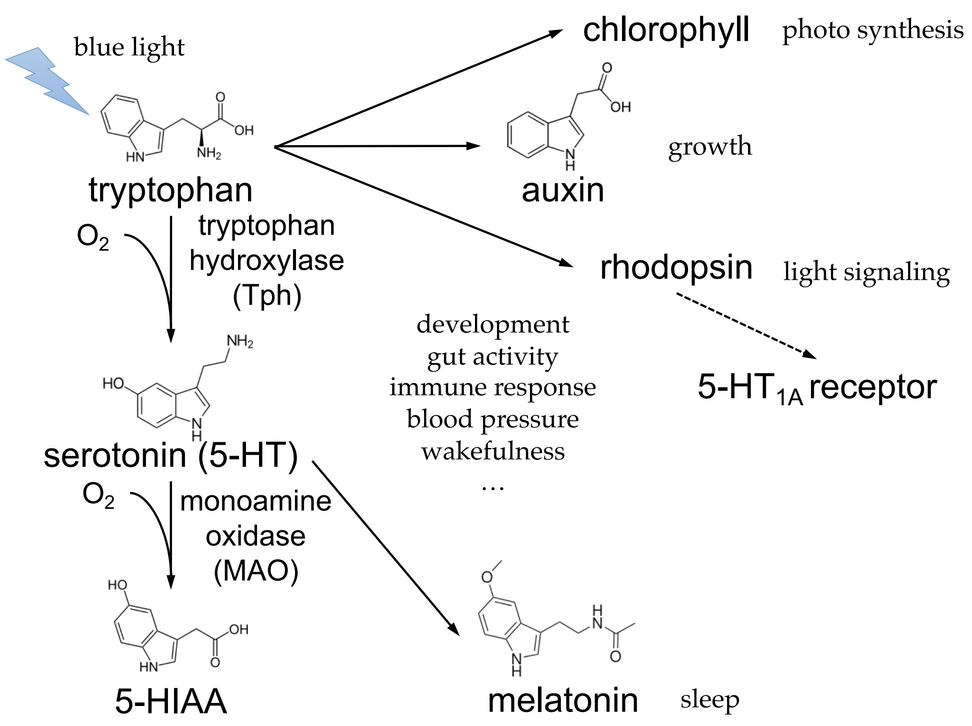
- They may share same evolutionary origin, but may have customized their messages for the audiences.

Serotonin Signals Available Time and Resources?

Serotonergic modulation of cognitive computations

Kenji Doya, Kayoko W Miyazaki and Katsuhiko Miyazaki

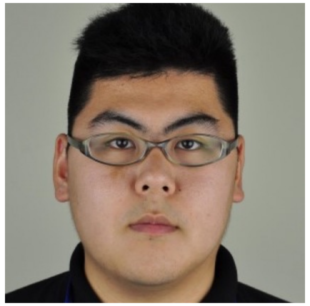
Current Opinion in Behavioral Sciences



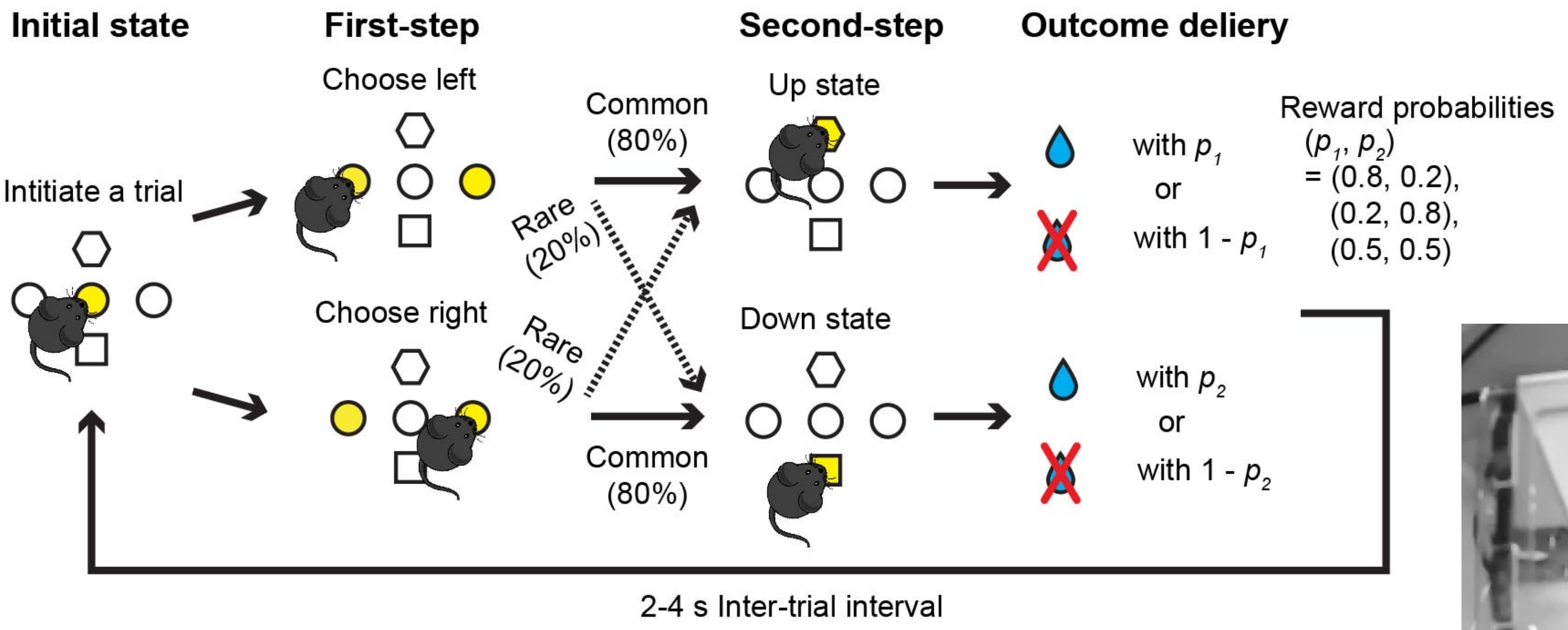
	Less time	More time
Development	stay	grow
Energy metabolism	utilize	save
Action vigor	spurt	relax
Risk taking	gamble	safe
Threat response	freeze, panic	cope, avoid
Social decision	selfish	cooperative
Learning rate α	fast	slow
Exploration β	exploit	explore
Temporal discounting γ	steep	slow
Eligibility 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

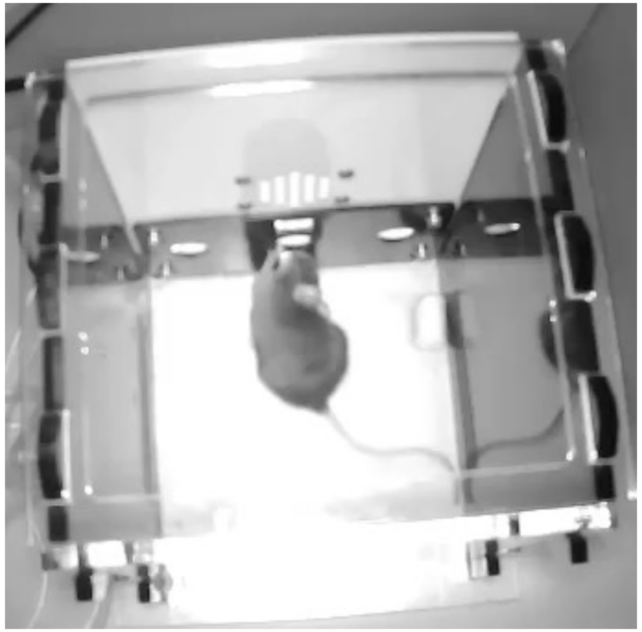
Serotonin for Model-based RL?

Masakazu Taira (COSYNE 2022)



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





Model-free/Model-based Hybrid RL Model

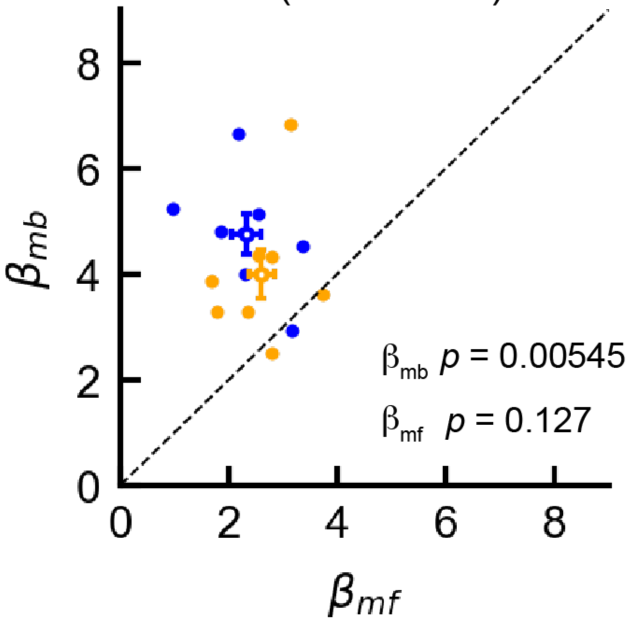
$$Q_{net}(a) = \beta_{\mathrm{mf}} Q_{mf}(a) + \beta_{mb} Q_{mb}(a)$$

$$P(a = \mathrm{left}) = \frac{1}{1 + \exp\left(-(Q_{net}(a = \mathrm{left}) - Q_{net}(a = \mathrm{right}) + P\bar{c} + B)\right)}$$

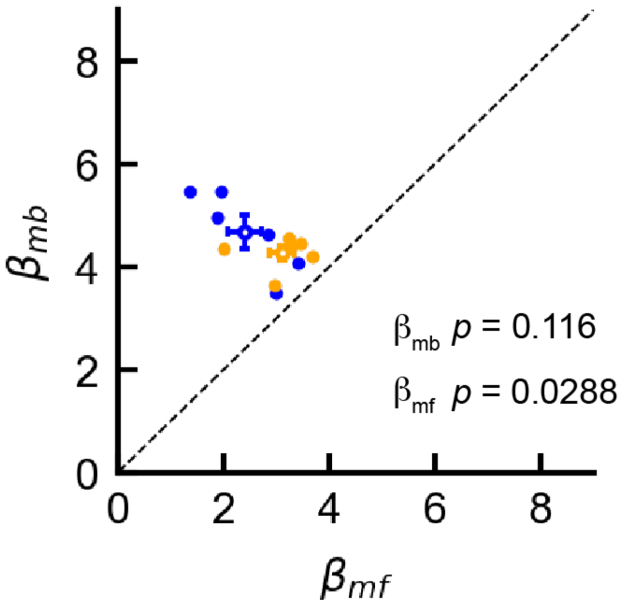
β_{mf}: weight for model-free strategy (0< β_{mf})
β_{mb}: weight for model-based strategy (0< β_{mb})
P: strength of choice perseveration
B: Choice bias

α: learning rate (0<α<1)
f: forgetting rate (0<f<1)
λ: eligibility trace (0< λ <1)

ArchT mice
(n = 8 mice)



Control mice
(n = 6 mice)



OIST Neural Computation Unit

Robotics/Machine Learning
- Christopher Buckley
- Shoko Ohta
- Qiong Huang
- Ho Ching Chiu
- Kristine Roque

Neural Modeling
- Carlos Gutierrez
- Sergio Flores
- Yukako Yamane
- Florian Larande
- Zhigang Mu

Neurobiology
- Katsuhiko Miyazaki
- Kayoko W Miyazaki
- Yuzhe Li
- Sergey Zobnin
- Masakazu Taira
- Miles Desforges

Administration
- Emiko Asato
- Kikuko Matsuo
- Misuzu Saito

Professor
- Kenji Doya

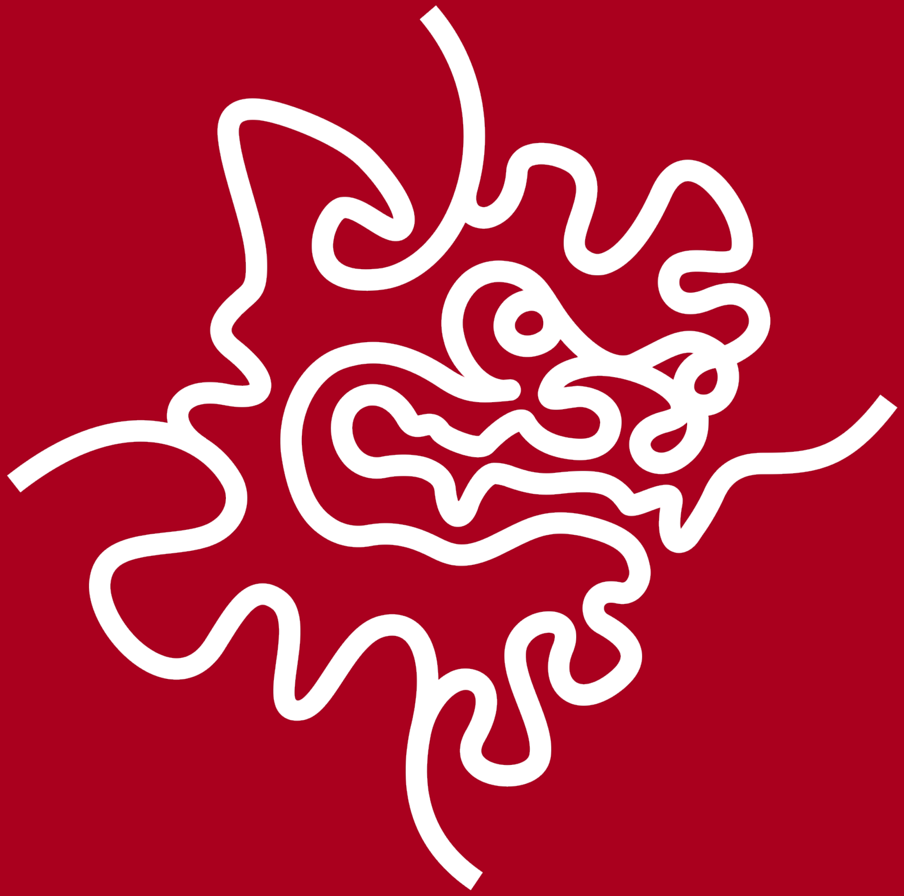
Acknowledgements

- 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**
  - **Thomas Akam (Oxford U)**
  - **Mark Walton (Oxford U)**
  - **Kenji Tanaka (Keio U)**
  - **Akihiro Yamanaka (Nagoya U)**

Scientific Research on Innovative Areas
Strategic Research Program for Brain Sciences

- Cortical imaging
  - **Akihiro Funamizu (U Tokyo)**
  - **Bernd Kuhn**
  - **Yuzhe Li**
  - **Sergey Zobnin**
- Marmoset data analysis
  - Carlos Gutierrez
  - Hiromichi Tsukada
  - Junichi Hata, Alex Woodward (RIKEN)
  - Ken Nakae, Henrik Skibbe (Kyoto U)
- Spiking neural network model
  - Jun Igarashi, Sun Zhe (RIKEN)
  - Tadashi Yamazaki, Hiroshi Yamamura (UEC)
  - Jan Moren
  - Osamu Shouno (HRI)
  - Benoit Girard, Daphne Heraiz (UPMC)
  - Jean Lienard
- Robotics
  - Jun Morimoto (ATR)
  - **Eiji Uchibe (ATR)**
  - Stefan Elfwing (ATR)
  - Jiexin Wang (ATR)
  - **Paavo Parmas (Kyoto U)**
  - Kristine Roque
  - Christopher Buckley

Brain/MINDS Project
Post-K Supercomputing Program



Thank you!