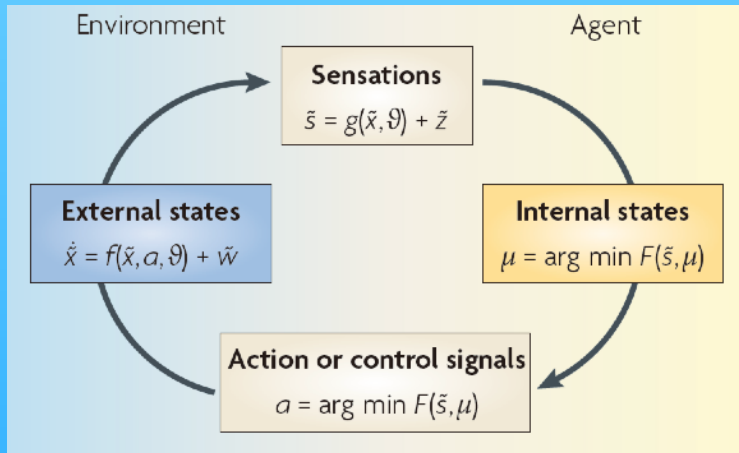


Active Inference

The Free Energy Principle in Mind, Brain, and Behavior



A recent theory on the function of the brain

perception
action

surprise
belief

learning
planning

generative model

Markov blanket

information entropy

variational free energy

KL divergence

expected free energy

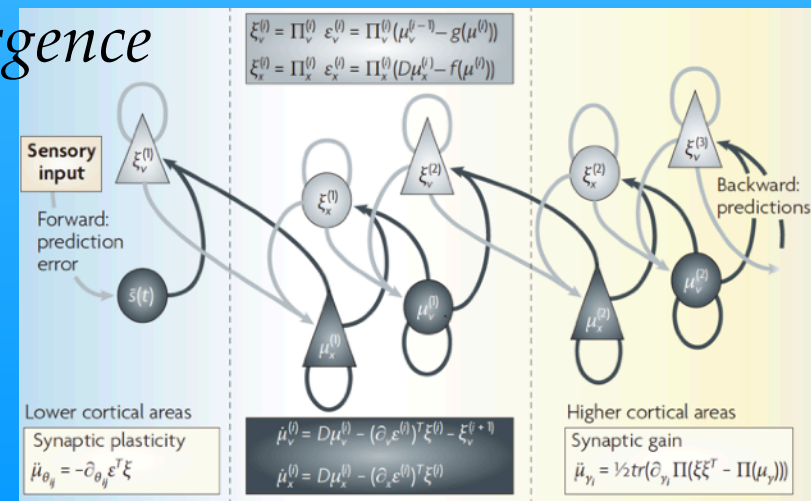
Karl Friston, *The free-energy principle: a unified brain theory?*

[Nature Reviews Neuroscience 11, 127 \(2010\)](#)

T. Parr, G. Pezzulo, and K. J. Friston, *Active Inference*, MIT, 2022

A. Tamii

RCNP, Univ. Osaka



Personal Motivations

Personal Motivation

Two years ago, I picked up *transformer* as the subject of my colloquium.

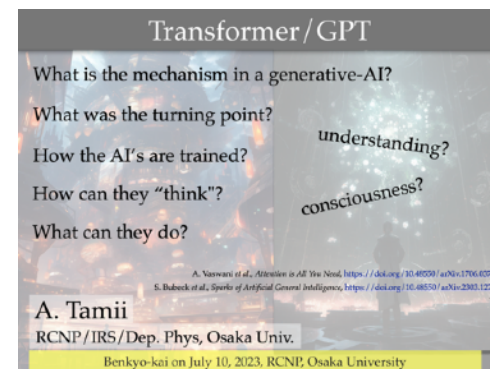
GPT = Generative Pre-trained Transformer

Generative-AI

But I didn't dig into the word *generative*.

The word *generative* originates from the word *generative model* in brain science.

Today, I will introduce a recent theory on the function of the brain, using the concepts of *active inference* and *free energy principle*, in relation to the word *generative model*.



Personal Motivation

I had many questions related the function of the brain and human behavior, *e.g.*

- how the nature produced, with the natural selection, a creature that can *understand* the laws of the universe
- observation of a brain wave 0.2 sec prior than one thought that one made a decision [5].
- why it is so difficult to walk up the steps of a stopped escalator.
- why a baby looks happy when he moves his/her hands or legs.
- how cognitive dissonance (認知的不協和) and self-justification (自己正当化) in behavioral psychology works
- function of excises and trainings for playing sports or musical instrumentations.
- patients who are unable to recognize what they see or the existence of the world on their left side [5].
- how memory modification happens

The answer or hints to the answer are given in the today's topic.

Free Energy Principle

“Any self-organizing system that is at equilibrium with its environment must minimize its free energy” [Friston06]

Karl J. Friston

The brain minimizes the free energy through inference for perception, action, and learning, resisting a tendency to disorder brought by the entropy law.

Unconscious Inference

neuroscience

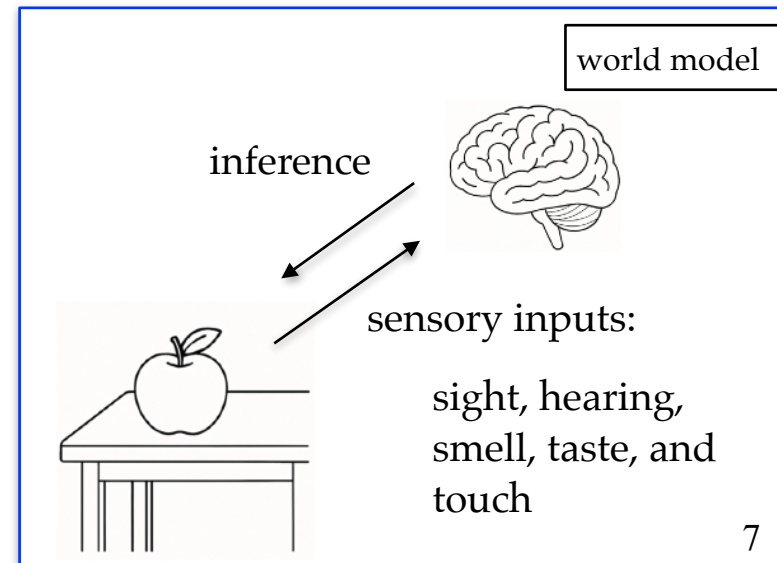
perception (知覚) as unconscious inference (無意識的推論) (Helmholtz 1867)

- The brain must **infer** the external world, e.g. the 3D structure, from the limited sensory data such as the visual input by the eyes.
- **The reality we perceive is not the world itself, but in the model of the world** that our brain has inferred.
- This process is **automatic, rapid, and unconscious**, drawing upon prior experiences and implicit knowledge.
- Perception operates much like a **probabilistic inference engine**, constantly generating hypotheses about the causes of sensory inputs (感覚).

Moving the eyes or body
→ the apple looks not moving.

Two eyes
→ one apple is perceived.

Immanuel Kant, *Critique of Pure Reason* (1781/1787):
“We can never know things in themselves, but only
appearances structured by our own faculties.”



Unconscious Inference

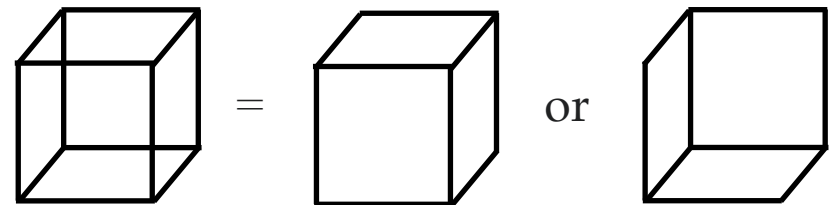
theoretical neuroscience



My Wife and My Mother-In-Law by W.E.Hill

After recognition of the illustration as a young woman, it is hard to recognize it as an old woman (vice versa).

The brain perceives the vision by *inference*.



Markov Blanket

theoretical neuroscience

The **agent** communicates with the **environment** only through the limited variables (*Markov Blanket*): sensory inputs (y) and actions (a).

Environment:

Hidden states (x) and causes (v)

→ creates the sensory inputs

generative process

Agent:

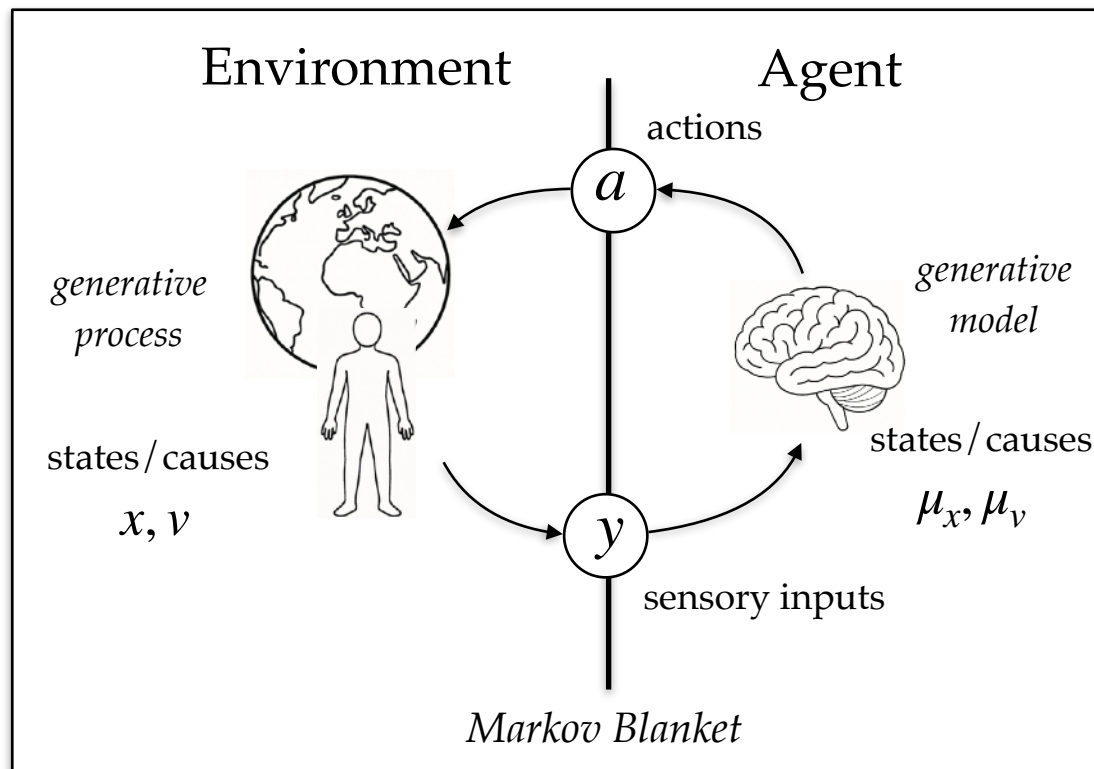
Parameters of hidden states μ_x
and causes (μ_v)

in a *generative model*

→ **infer** the sensory inputs.

The agent can change the environment
(and sensory inputs) through the action.

Markov Blankets can be hierarchical.



The environment must be orderly structured
and is logically understandable.

Hidden States and Hidden Causes

theoretical neuroscience

Hidden states: more temporal states, changes more quickly than causes

e.g.

the color of the surface of a metal box

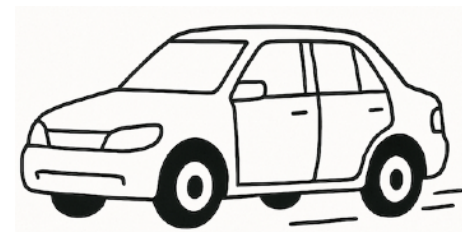
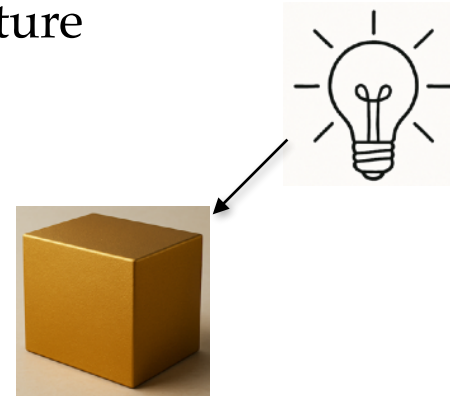
speed of a car

Hidden causes: more regular and fundamental like laws in nature

e.g.

property of metal

equation of motion

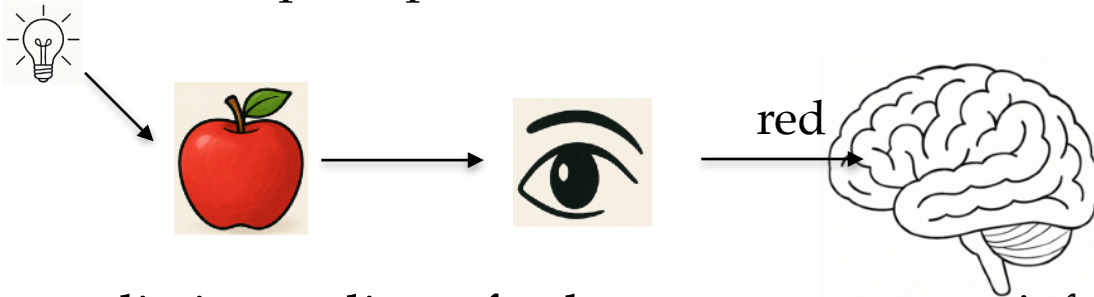


Active Inference

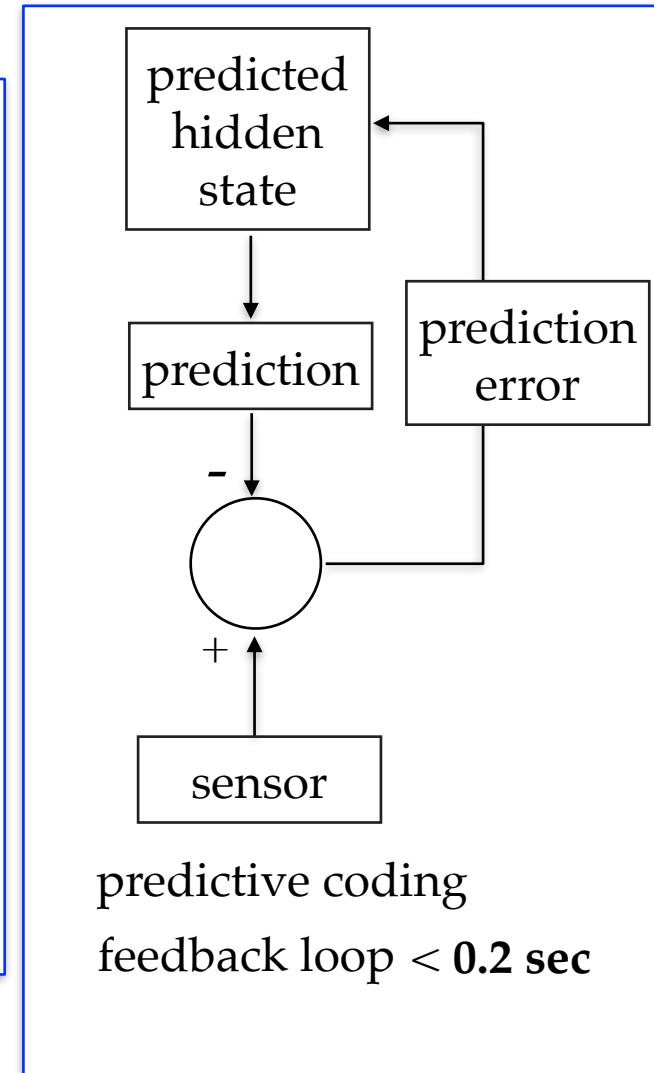
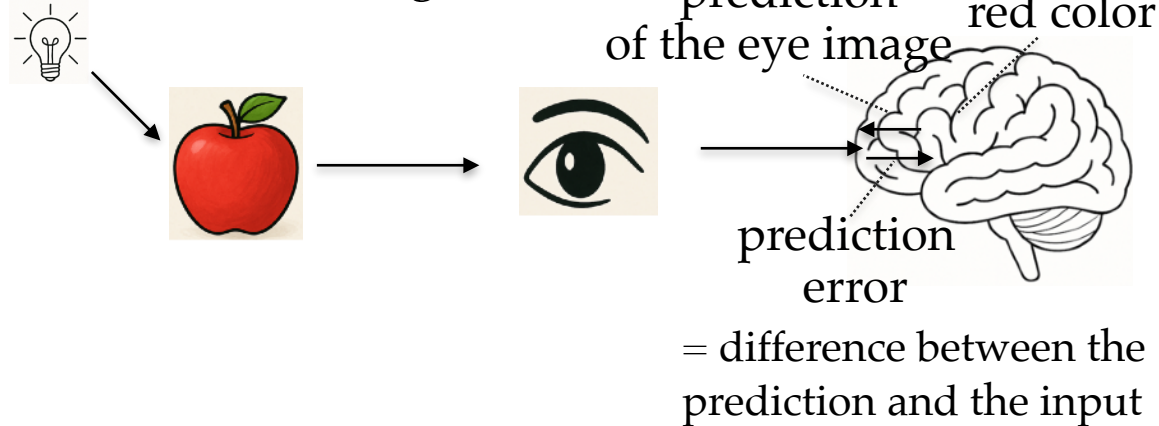
Predictive Coding

Coding = conversion of the inputs to an internal expression

traditional perception of color



predictive coding of color



The color can be recognized under different type of lights.
The color is predicted according to the situation.

p.d.f. is inferred by a brain

What is inferred by a brain is **not the value** itself of a hidden state x (or a hidden causes v) **but the probability density distribution** (p.d.f.) (確率密度分布) of it.

$P(x)$: a p.d.f. of x

$$P(x) \geq 0 \quad \int dx P(x) = 1$$

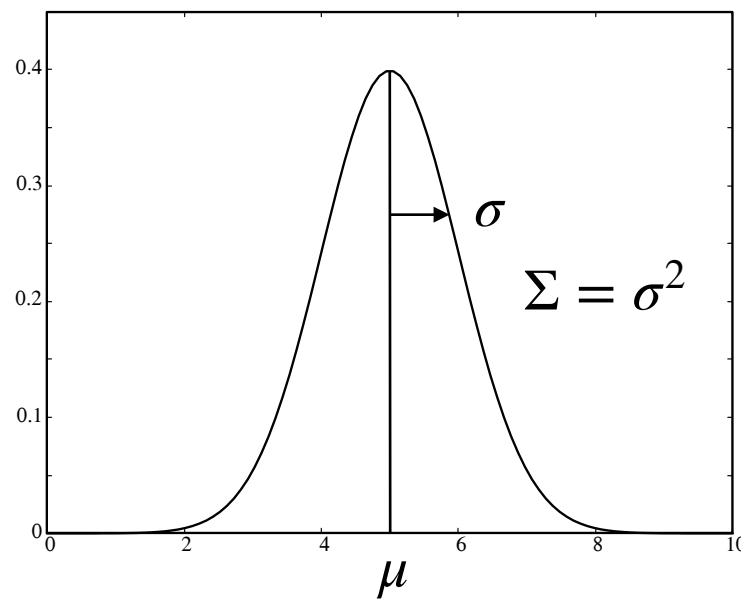
In a simplified representation using Gaussian approximation, μ and Σ (or Π) are the model parameters to be inferred.

μ : mean (平均值)

Σ : variance (分散)

$\Pi \equiv \Sigma^{-1}$: precision (精度)

Gaussian approximation of a p.d.f.



The same for discrete (category) states, e.g. an apple 80%, a pear 15%, ...

Bayesian Inference

statistics

$P(x)$: probability density function (p.d.f.) $\int dx P(x) = 1$

$P(x, y)$: joint p.d.f.

$P(x|y)$: conditional p.d.f. p.d.f. of x when y

Probabilistic reasoning (Bayesian reasoning)

$$P(x, y) = P(y|x)P(x) = P(x|y)P(y)$$

$$P(x|y) = \frac{P(y|x)P(x)}{P(y)}$$

$P(x)$: prior p.d.f. (事前確率分布) prior belief (事前信念)

$P(x|y)$: posterior p.d.f. (事後確率分布) posterior belief (事後信念)
= updated p.d.f. after observing y

$$\int dx P(x, y) = \int dx P(y|x)P(x) = P(y) \int dx P(x|y) = P(y)$$

$P(y)$: marginal probability (周辺確率)

for today's topic
 x : hidden state
 y : sensory inputs

in perception, x : internal hidden state, y : sensory inputs

Shannon Information Content

information theory

Claude Shannon (1948)

Information content (情報量): $I(x)$

when observing an event x

$$\begin{aligned} I(x) &= -\log_2 P(x) && \text{in the unit of bits} \\ &= -\ln P(x) && \text{in the unit of nats} \end{aligned}$$

observation of a rarer event \rightarrow getting a larger information content

Information Entropy (Shannon Entropy) (情報エントロピー) : H

$$H(p(x)) = - \int dx P(x) \ln P(x)$$

averaged (expected) information content

larger entropy \rightarrow larger uncertainty, larger randomness

smaller entropy \rightarrow more predictability, more order

Surprise

In the context of the free energy principle

$-\ln P(y)$: *surprise*

rarer event observation \rightarrow larger surprise

$$P(y) = \int dx P(x, y) = \int dx P(y|x)P(x) = P(y) \int dx P(x|y)$$

$P(y)$: marginal likelihood

$P(x, y)$: generative model

Variational Free Energy: Definition

theoretical neuroscience

Surprise (or negative log model evidence) is written for any p.d.f. $Q(x)$ (>0 for any x)

$$-\ln P(y) = -\ln \int dx P(x, y) \frac{Q(x)}{Q(x)}$$

y : sensory inputs

x : hidden state

$$= -\ln \mathcal{E}_{Q(x)} \left[\frac{P(x, y)}{Q(x)} \right]$$

$\mathcal{E}_{g(x)} [f(x)]$: expectation of $f(x)$ with a weight function of $g(x)$.

$$\leq -\mathcal{E}_{Q(x)} \left[\ln \frac{P(x, y)}{Q(x)} \right]$$

$$\equiv \int dx g(x) f(x)$$

$$\equiv F(Q, y)$$

\therefore Jensen's inequality

equality holds when $P(x, y) = Q(x)$

$F [Q, y]$: Variational Free Energy

$$F(Q, y) \equiv -\mathcal{E}_{Q(x)} \left[\ln \frac{P(x, y)}{Q(x)} \right] = -\mathcal{E}_{Q(x)} [\ln P(x, y)] + \mathcal{E}_{Q(x)} [\ln Q(x)]$$

Variational free energy is the **upper-bound of the surprise**.

Variational Free Energy and Surprise

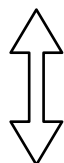
$$P(x, y) = P(x | y)P(y)$$

Practically, it is difficult to calculate exactly $P(x | y)$.

$$\ln P(x, y) = \ln P(x | y) + \ln P(y)$$

$$\mathcal{E}_{P(x|y)} [\ln P(x, y)] = \mathcal{E}_{P(x|y)} [\ln P(x | y)] - (-\ln P(y))$$

expectation by x with the weight of $P(x | y)$



comparison

$$\mathcal{E}_{Q(x)} [\ln P(x, y)] = \mathcal{E}_{Q(x)} [\ln Q(x)] - F(Q, y)$$

from the previous slide

$Q(x)$ is the **approximated posterior probability** $P(x | y)$,
called *recognition density* (認識分布)

If the approximation is perfect (imperfect), the variational free energy coincides with (is the upper-bound of) the surprise.

Minimization of the variational free energy

= minimization of surprise & better approximation of $P(x | y)$

better knowledge less surprise

Free Energy Principle

“Any self-organizing system that is at equilibrium with its environment must minimize its free energy” [Friston06]

The brain minimizes the free energy through inference for perception, action, and learning, resisting a tendency to disorder and the entropy law.

Kullback-Leibler (KL) Divergence: D_{KL}

A measure of the difference between the two p.d.f.'s: $q(x)$ and $p(x)$

$$D_{KL}(q(x) || p(x)) \equiv \int_{-\infty}^{+\infty} dx q(x) \ln \frac{q(x)}{p(x)}$$

Note that the D_{KL} is not symmetric between $q(x)$ and $p(x)$.

$$D_{KL}(q(x) || p(x)) \geq 0 \quad \text{Equality holds when } q(x) = p(x).$$

$$\begin{aligned} D_{KL}(q(x) || p(x)) &= \mathcal{E}_{q(x)} \left[\ln \frac{q(x)}{p(x)} \right] \\ &= \mathcal{E}_{q(x)} [\ln q(x) - \ln p(x)] \\ &= \mathcal{E}_{q(x)} [\ln q(x)] - \mathcal{E}_{q(x)} [\ln p(x)] \\ &= -H(q(x)) - \mathcal{E}_{q(x)} [\ln p(x)] \end{aligned}$$

Variational Free Energy: definition with D_{KL}

$$\begin{aligned} F(Q, y) &= - \mathcal{E}_{Q(x)} \left[\ln \frac{P(x, y)}{Q(x)} \right] \\ &= - \mathcal{E}_{Q(x)} \left[\ln \frac{P(x | y) P(y)}{Q(x)} \right] \\ &= - \mathcal{E}_{Q(x)} \left[\ln \frac{P(x | y)}{Q(x)} \right] - \mathcal{E}_{Q(x)} [\ln P(y)] \\ &= \mathcal{E}_{Q(x)} \left[\ln \frac{Q(x)}{P(x | y)} \right] - \ln P(y) \\ &= \underbrace{D_{KL} (Q(x) || P(x | y))}_{\text{Divergence}} - \underbrace{\ln P(y)}_{\text{Log Model Evidence}} \end{aligned}$$

Internal Energy

Internal Energy $U(x; y)$ is defined as

$$U(x; y) \equiv -\ln p(x, y) = -\ln p(y|x) - \ln p(x)$$

Then

$$\begin{aligned} F(Q, y) &= -\mathcal{E}_{Q(x)} \left[\ln \frac{P(x, y)}{Q(x)} \right] \\ &= -\mathcal{E}_{Q(x)} [\ln P(x, y)] + \mathcal{E}_{Q(x)} [\ln Q(x)] \\ &= \underbrace{\int dx Q(x) U(x; y)}_{\text{posterior expectation of the internal energy (w.r.t. the recognition density)}} - \underbrace{H(Q(x))}_{\text{entropy of the recognition density}} \end{aligned}$$

c.f. in statistical mechanics

$$p(x) \propto \exp \left(-\frac{U(x)}{k_B T} \right)$$

: Boltzmann factor

$$F = U - TS$$

: Helmholtz free energy

Analogy to the Thermodynamics

Helmholtz Free Energy in Thermodynamics

under isothermal ($T=\text{const.}$) and isochoric ($V=\text{const.}$) conditions

$$F = U - TS$$

represents the maximum work that is extractable from the system

$$F [Q, y] = \underbrace{- \mathcal{E}_{Q(x)} [\ln P(x, y)]}_{\text{Energy}} - \underbrace{H [Q(x)]}_{\text{Entropy}}$$

$P(x, y)$ is the generative model that (the log of it) is analogous to the internal energy.

The system evolves to the direction of decreasing the Helmholtz free energy.
(second law of thermodynamics, the entropy low)

Variational Free Energy: Reformulations

Variational Free Energy

$$F [Q, y] = \underbrace{- \mathcal{E}_{Q(x)} [\ln P(x, y)]}_{\text{Energy}} - \underbrace{H [Q(x)]}_{\text{Entropy}}$$

$$= \underbrace{D_{KL} [Q(x) || P(x)]}_{\text{Complexity}} - \underbrace{\mathcal{E}_{Q(x)} [\ln P(y | x)]}_{\text{Accuracy}}$$

$$= \underbrace{D_{KL} [Q(x) || P(x | y)]}_{\text{Divergence}} - \underbrace{\ln P(y)}_{\text{Evidence}}$$

Generalized Coordinate of Hidden States

Analytical Mechanics

a set of hidden states $\tilde{\mathbf{x}} = (\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n) = \mathbf{x}_i$

each \mathbf{x}_i is a vector as

$$\mathbf{x}_i = \begin{bmatrix} x_i \\ x'_i \\ x''_i \\ \vdots \end{bmatrix}$$

x_i : “location” of a hidden state in the generalized coordinate

x'_i : “velocity”

x''_i : “acceleration”

\vdots

The words “location” etc. are not of the exact meanings but are analogical concepts.

Operator D generates a vector of time derivative.

$$D\mathbf{x}_i = \begin{bmatrix} x'_i \\ x''_i \\ x'''_i \\ \vdots \end{bmatrix} = \begin{bmatrix} 0, 0, 0, \dots \\ 1, 0, 0, \dots \\ 0, 1, 0, \dots \\ \vdots \end{bmatrix} \begin{bmatrix} x_i \\ x'_i \\ x''_i \\ \vdots \end{bmatrix}$$

Free Energy Minimization

Assume the normal distribution of $Q(\tilde{\mathbf{x}})$: *Laplace Assumption*

Then, the equation

$$F(Q, y) = \int d\tilde{\mathbf{x}} Q(\tilde{\mathbf{x}}) U(\tilde{\mathbf{x}}; y) - H(Q(\tilde{\mathbf{x}}))$$

is rewritten as

$$F(\tilde{\boldsymbol{\mu}}, \Sigma; y) = U(\tilde{\boldsymbol{\mu}}; y) + \frac{1}{2} \text{tr} (\Sigma \nabla^2 U(\tilde{\boldsymbol{\mu}}; y)) - \frac{1}{2} \ln |\Sigma| - \frac{n}{2} \ln(2\pi e) \quad \text{see [3]}$$

$$[\nabla^2]_{ij} \equiv \frac{\partial^2}{\partial x_i \partial x_j}$$

The condition of having minimum

$$\frac{\partial F(\tilde{\boldsymbol{\mu}}, \Sigma; y)}{\partial \mu_i} = 0 \quad \Rightarrow \quad \Sigma = (\nabla^2 U(\tilde{\boldsymbol{\mu}}; y))^{-1}$$

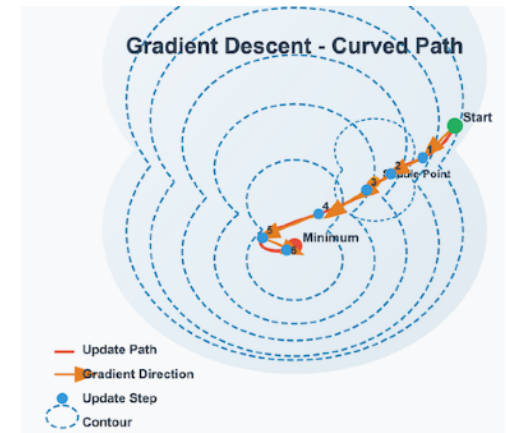
Iterative minimization by gradient descent (勾配降下法)

$$\frac{d}{dt} \mu_i = - \frac{\partial F(\tilde{\boldsymbol{\mu}}, \Sigma; y)}{\partial \mu_i} \rightarrow 0$$

μ : mean

Σ : covariant tensor
(variance for a variable)

$\Pi \equiv \Sigma^{-1}$: precision



similar to numerical solution
of simultaneous equations

Model Evidence, Sensory Entropy

More explicitly, $P(y)$ depends on the generative model m , thus denoted as $P(y | m)$.

$P(y | m)$ is interpreted as evidence of the model m .

= measure of the goodness of the model

$\ln P(y | m)$ is called *log model evidence* (= negative surprise)

The leaning process is formulated as the maximization of the evidence, or equivalently with the **minimization** of the *sensory entropy* $H(y | m)$.

$$\begin{aligned} H(y | m) &= - \int dy P(y | m) \ln P(y | m) && \text{average over all } y \\ &= - \lim_{T \rightarrow \infty} \frac{1}{T} \int_0^T dt \ln P(y(t) | m) && \begin{array}{c} \downarrow \text{ergodicity} \\ \text{average over a long time} \end{array} \end{aligned}$$

Instead of minimizing the sensory entropy, its upper bound, the time integration of the Free energy

$$S(y, q) \equiv \int_0^t F(\tau; y, q) d\tau \quad \text{variational action of the free energy}$$

is minimized.

Free Energy Minimization

In a time-dependent dynamic system, the variational action is minimized in the trajectory of the mean parameters:

$$\dot{\mu}_i - D\mu_i = - \frac{\partial S(\tilde{\mu}, \Sigma; y)}{\partial \mu_i}$$

The condition of having minimum

$$\frac{\partial S(\tilde{\mu}, \Sigma; y)}{\partial \mu_i} = 0 \quad \Rightarrow \quad \dot{\mu}_i = D\mu_i$$

in analytical mechanics

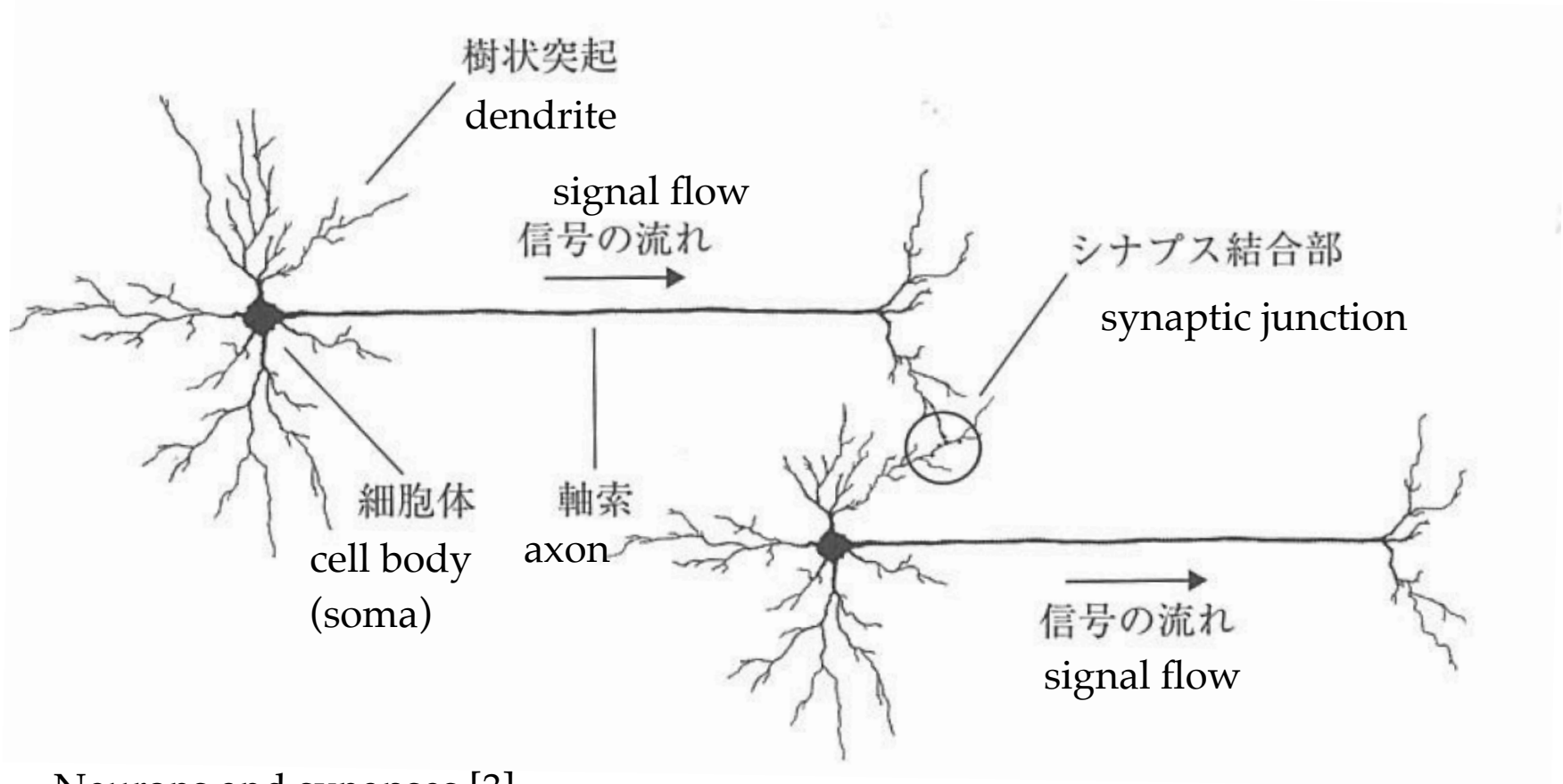
$$S[\mathbf{q}(t)] = \int_{t_1}^{t_2} L(\mathbf{q}(t), \dot{\mathbf{q}}(t), t) dt$$

$$\frac{d}{dt} \left(\frac{\partial L}{\partial \dot{q}_i} \right) - \frac{\partial L}{\partial q_i} = 0$$

Function of Neurons in a Brain

Neurons and Synapses

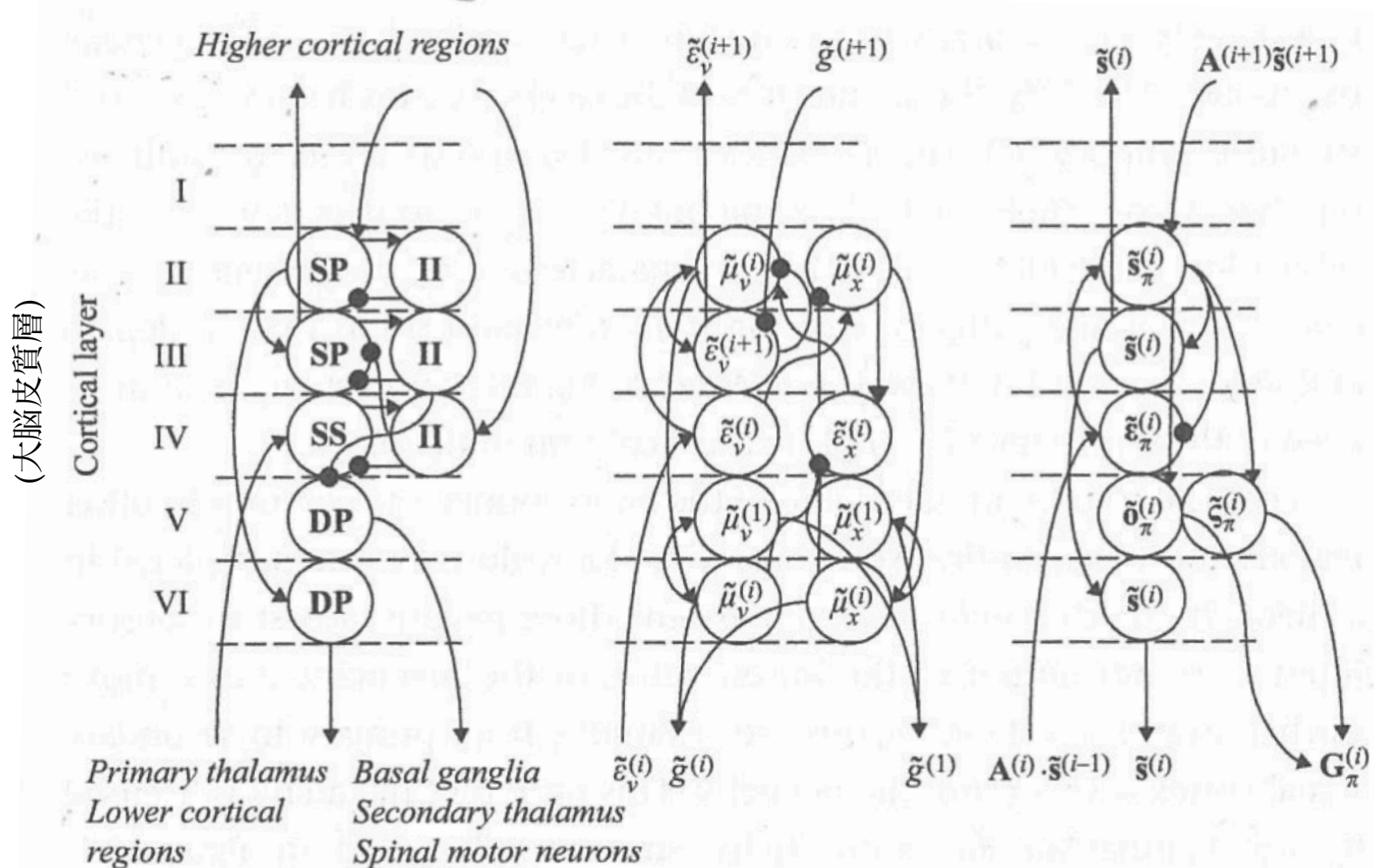
neurobiology



Neurons and synapses [3]

axon length: 1 μm ~1 cm in the brain connection

Neuron Cells



Hierarchical structure of the human brain [2]

Hierarchical Model

[Friston08]

bottom layer

$$y = g^{(1)}(x^{(1)}, v^{(1)}) + \epsilon_y^{(1)}$$

$$Dx^{(1)} = f^{(1)}(x^{(1)}, v^{(1)}) + \epsilon_x^{(1)}$$

i th layer

$$v^{(i-1)} = g^{(i)}(x^{(i)}, v^{(i)}) + \epsilon_v^{(i)}$$

$$Dx^{(i)} = f^{(i)}(x^{(i)}, v^{(i)}) + \epsilon_x^{(i)}$$

top layer ($n+1$ th)

$$v^{(n)} = \eta + \epsilon_v^{(n+1)}$$

x^i : hidden state (i th layer)

v^i : hidden cause (i th layer)

D : differentiation in time

$\epsilon_v^{(i)}$: noise in $v^{(i-1)}$

$\epsilon_x^{(i)}$: noise in $Dx^{(i)}$

$g^{(i)}$: *observer equation*

$f^{(i)}$: *state equation*

η : the top hidden cause
(randomly inferred)

Hierarchical Model

[Friston08]

Prediction errors

$$e \equiv \begin{bmatrix} e^x \\ e^v \end{bmatrix}$$

$$e^x \equiv \begin{bmatrix} Dx^{(1)} \\ Dx^{(2)} \\ \vdots \\ Dx^{(n)} \end{bmatrix} - \begin{bmatrix} f(x^{(1)}, v^{(1)}) \\ f(x^{(2)}, v^{(2)}) \\ \vdots \\ f(x^{(n)}, v^{(n)}) \end{bmatrix}$$

$$e^v \equiv \begin{bmatrix} y \\ v^{(1)} \\ \vdots \\ v^{(n)} \end{bmatrix} - \begin{bmatrix} g(x^{(1)}, v^{(1)}) \\ g(x^{(2)}, v^{(2)}) \\ \vdots \\ \eta \end{bmatrix}$$

Renormalized prediction errors

$$\xi \equiv \Pi e = \Sigma^{-1} e = e - \Lambda \xi$$

Σ : covariance matrix

$\Pi \equiv \Sigma^{-1}$: precision

$$\dot{\mu}_v^{(i)} = D\mu_v^{(i)} - \delta_v^{(i)T} \xi_v^{(i)} - \xi_v^{(i+1)}$$

$$\dot{\mu}_x^{(1)} = D\mu_x^{(1)} - \delta_x^{(1)T} \xi_x^{(1)}$$

$$\xi_v^{(i)} = \mu_v^{(i-1)} - g(\mu_v^{(i)}, \mu_x^{(i)}) - \Lambda_v^{(i)} \xi_v^{(i)}$$

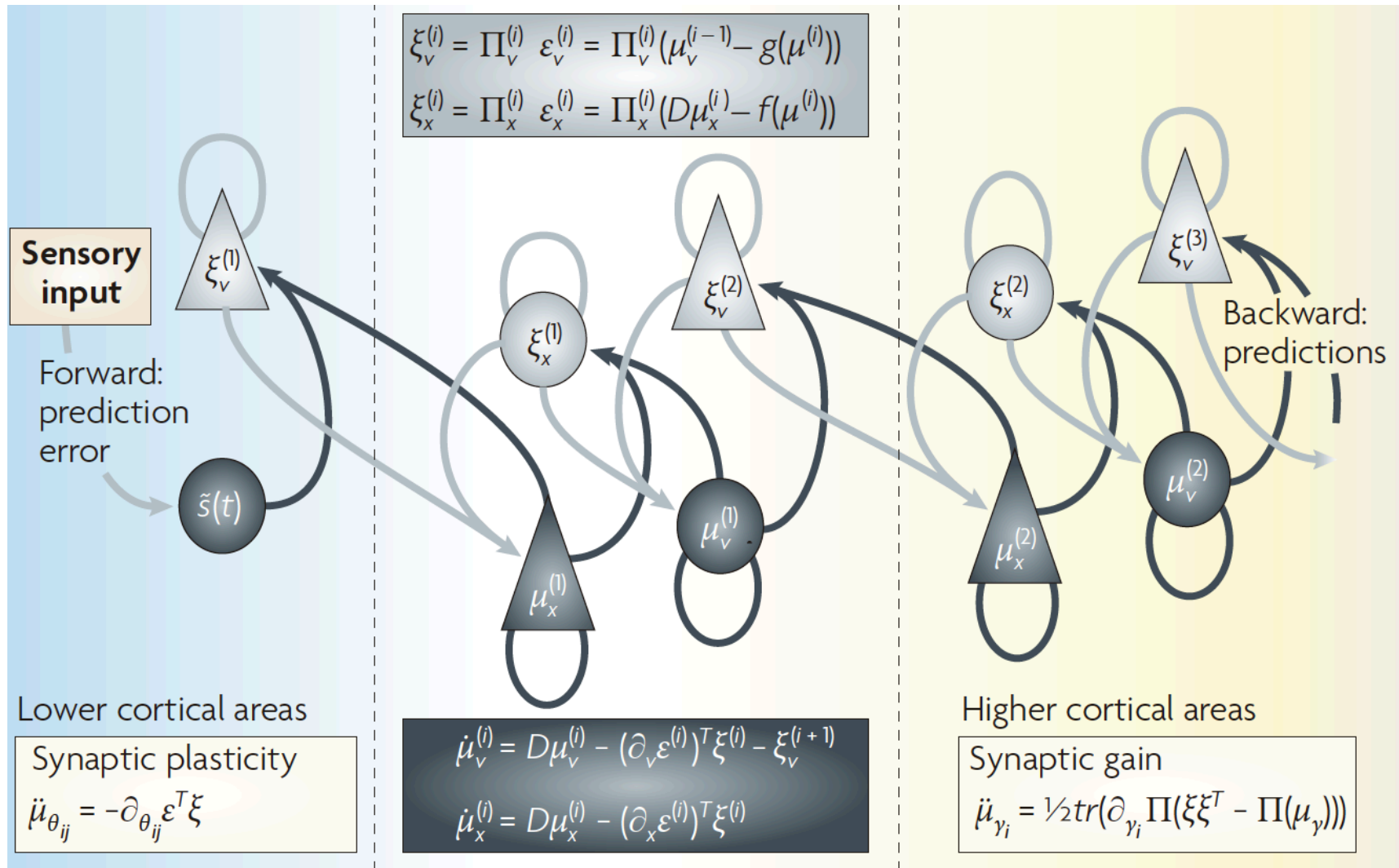
$$\xi_x^{(i)} = D\mu_x^{(i)} - f(\mu_v^{(i)}, \mu_x^{(i)}) - \Lambda_x^{(i)} \xi_x^{(i)}$$

$$\delta_x \equiv \frac{\partial e_x}{\partial \mu_x} \quad \delta_v \equiv \frac{\partial e_v}{\partial \mu_v}$$

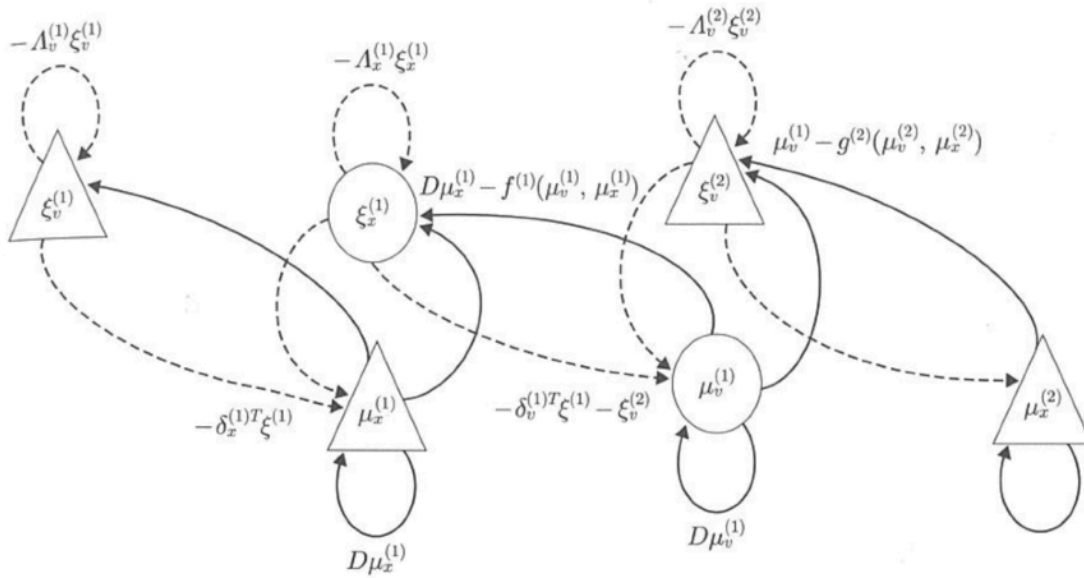
$$\dot{\mu} - D\mu = - \frac{\partial S(\tilde{\mu}, \Sigma; y)}{\partial \mu}$$

The generation model described by $Q(x)$ and $P(x, y)$ is implemented in a system with the functions of $f^{(i)}$ and $g^{(i)}$ and the parameters of $\mu_x^{(i)}$, $\mu_v^{(i)}$, $\xi_x^{(i)}$, and $\xi_v^{(i)}$.

Hierarchical Message Passing



Message Passing



message passing in a cortical layer [3]

$$\dot{\mu}_v^{(i)} = D\mu_v^{(i)} - \delta_v^{(i)T}\xi_v^{(i)} - \xi_v^{(i+1)}$$

$$\dot{\mu}_x^{(1)} = D\mu_x^{(1)} - \delta_x^{(1)T}\xi_x^{(1)}$$

$$\xi_v^{(i)} = \mu_v^{(i-1)} - g(\mu_v^{(i)}, \mu_x^{(i)}) - \Lambda_v^{(i)}\xi_v^{(i)}$$

$$\xi_x^{(i)} = D\mu_x^{(i)} - f(\mu_v^{(i)}, \mu_x^{(i)}) - \Lambda_x^{(i)}\xi_x^{(i)}$$

$$\delta_x \equiv \frac{\partial e_x}{\partial \mu_x} \quad \delta_v \equiv \frac{\partial e_v}{\partial \mu_v}$$

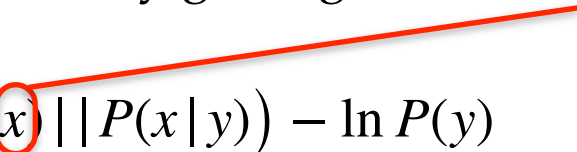
$$\dot{\mu} - D\mu = - \frac{\partial S(\tilde{\mu}, \Sigma; y)}{\partial \mu}$$

Perception, Action, Learning

Perception as Inference

- The 3D structure is predicted when a 2D image is watched.
- One listens to the talk of another one with predicting what will be told next.
- The outside scenery watched from a moving train looks stable.

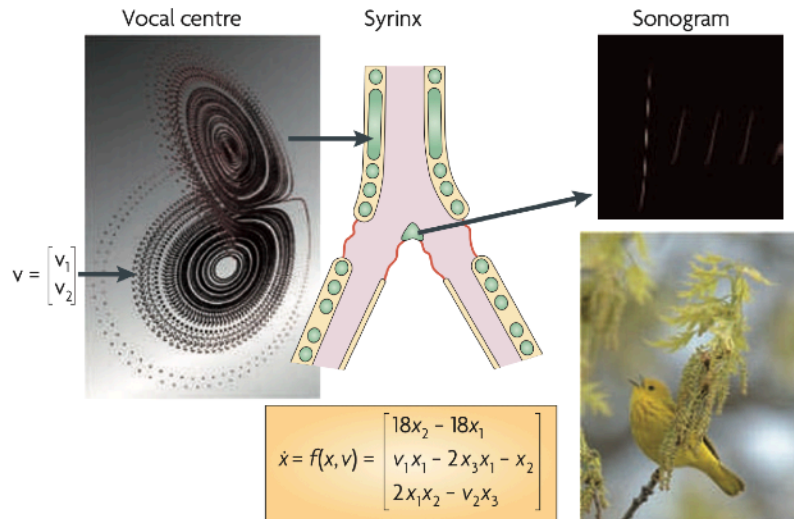
free energy minimization by getting better knowledge on the hidden states

$$F(Q, y) = \underbrace{D_{KL}(Q(x) || P(x|y))}_{\text{Divergence}} - \underbrace{\ln P(y)}_{\text{Evidence}}$$


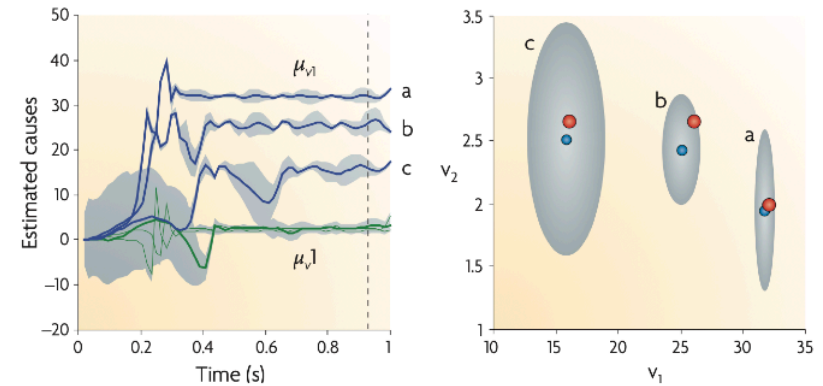
Perception as Inference

Simulation on birdsong perceptual categorization (Fig 1 in [1])

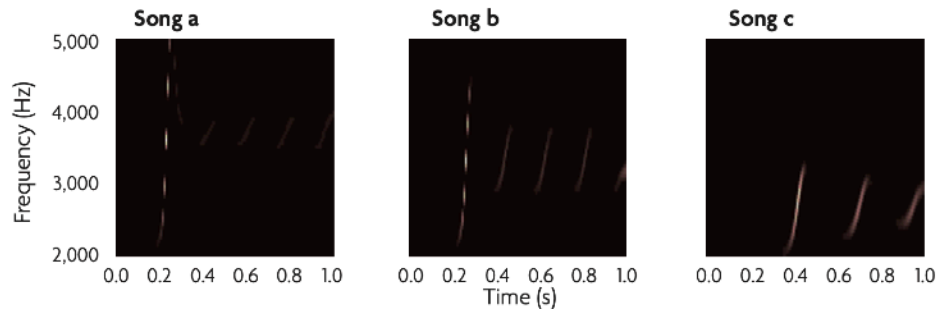
a Perceptual inference



c

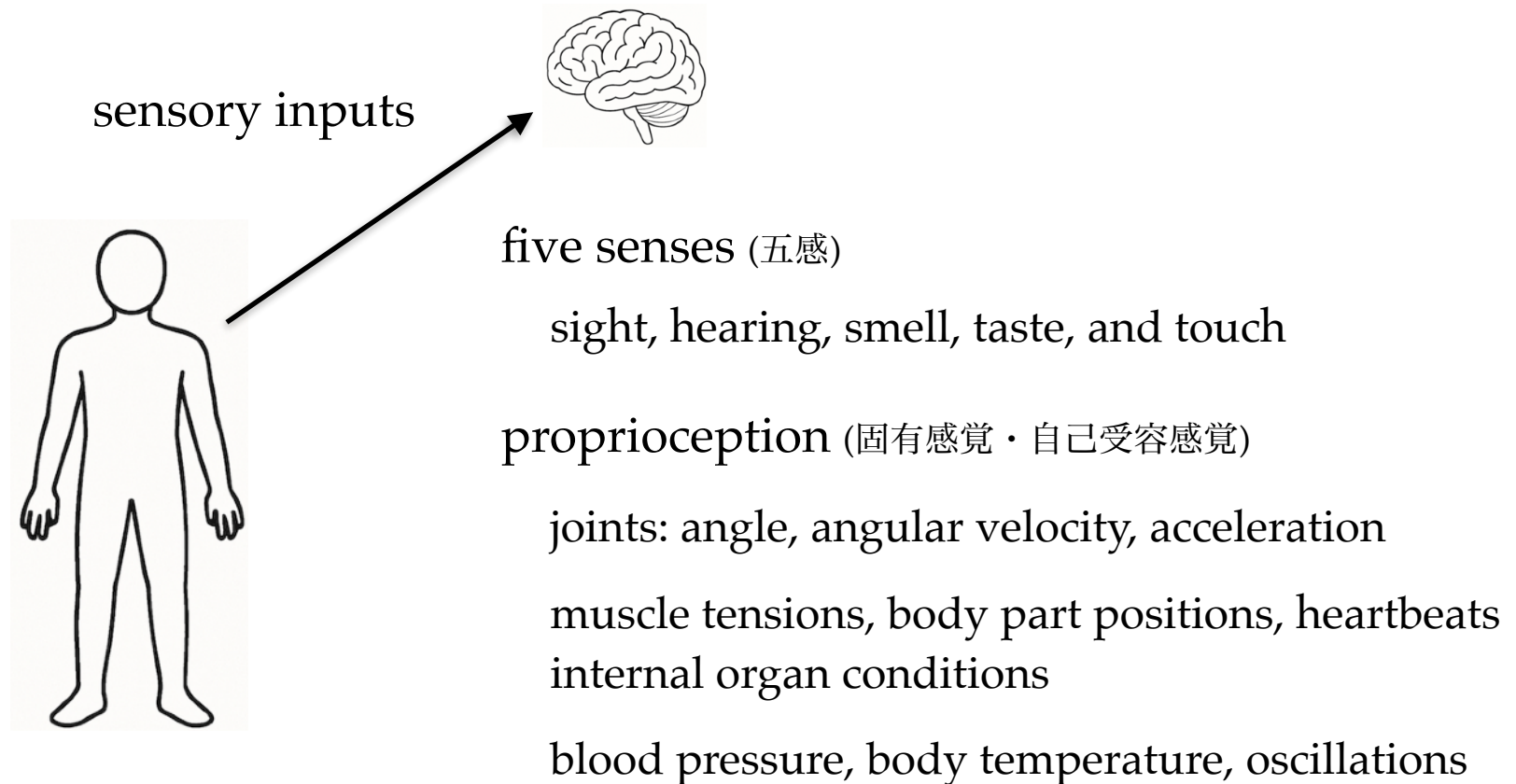


b Perceptual categorization



Proprioception

Sensory inputs include the signals coming from the body.



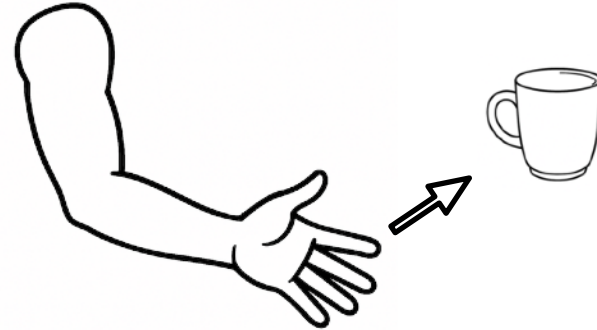
※Emotions are originated from the internal organ conditions.

Action by Inference

traditional control of action



control the muscles
(joint angles, hand position)



action by inference



predict (infer)
joint & muscle conditions
vision, touch, sound,..l
of the next moment



get sensory inputs
→ calculate the error (difference from the prediction)
→ feed back

free energy minimization by changing the environment

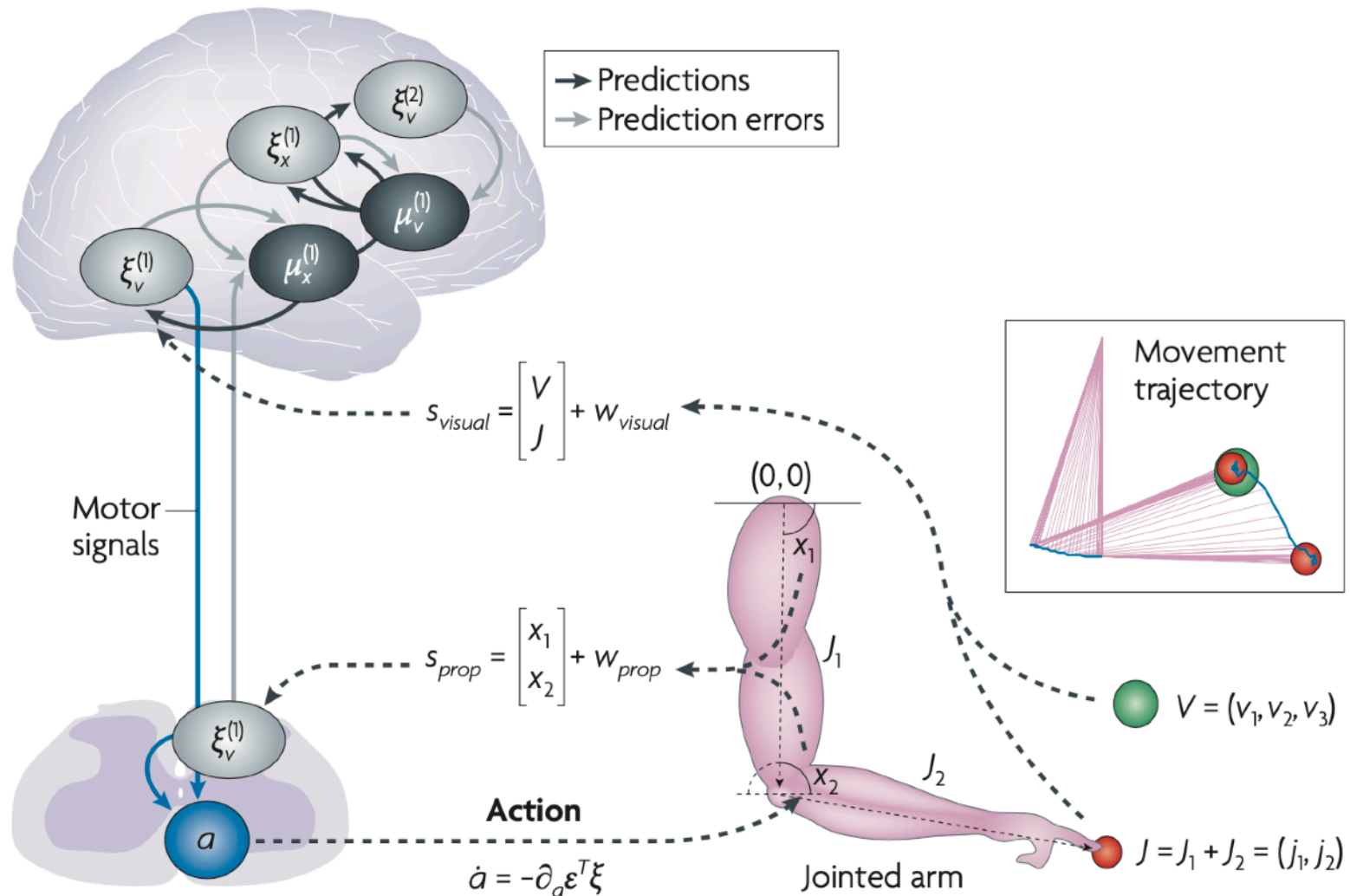
$$F(Q, y) = D_{KL} (Q(x) || P(x | y)) - \ln P(y)$$

Divergence Evidence

reducing surprise
increasing negative log evidence

Action

A demonstration of cued reaching movements (Fig 2 in [1])



Examples of Action as Inference

- *Homeostasis* (生物学的恒常性)
- *Sense of agency* (自己主体感)
- I feel difficulty to walk up the steps of a stopped escalator.
- ...

※ A baby feels happy with observing sense of agency.

Active Inference in Discrete Time

Partially Observable Markov Decision Processes (POMDP)

matrix representation

- 1 $P(\pi) \equiv \text{Cat}(\pi_0)$
- 2 $P(O_\tau | s_\tau) \equiv \text{Cat}(A)$
- 3 $P(s_{\tau+1} | s_\tau, \pi) \equiv \text{Cat}(B_{\pi\tau})$

$$P(O_\tau = i | s_\tau = j) \equiv A_{ij}$$

$$P(s_1) \equiv \text{Cat}(D)$$

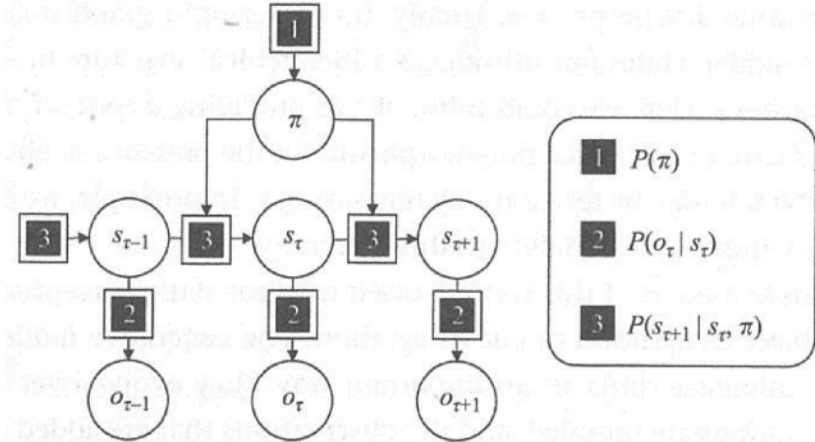
$$P(\tilde{s} | \pi) \equiv P(s_1) \prod_{\tau=1} P(s_{\tau+1} | s_\tau, \pi) = \text{Cat}(D) \prod_{\tau=1} \text{Cat}(B_{\pi\tau})$$

$$\pi_0 = \sigma(-G) \quad \sigma: \text{softmax function for normalization}$$

C : parameters in the generative model

$$G_\pi = G(\pi) = -\mathcal{E}_{\tilde{Q}} \left[D_{KL} \left[Q(\tilde{s} | \tilde{o}, \pi) || Q(\tilde{s} | \pi) \right] \right] - \mathcal{E}_{\tilde{Q}} \left[D_{KL} \ln P(\tilde{o} | C) \right] \quad \text{: expected free energy}$$

$$\tilde{Q}(o_\tau, s_\tau | \pi) \equiv P(o_\tau | s_\tau) Q(s_\tau | \pi) \quad \text{: posterior predictive density}$$



τ : time step

o_τ : outcome

s_τ : hidden state

π : policy

\tilde{s} : time sequence of s_τ

Expected Free Energy

$$G(\tau; \pi_i) \simeq \underbrace{\mathcal{E}_q [\ln q(o | \pi_i) - \ln q(o | s, \pi_i)]}_{\substack{\text{(-) epistemic value} \\ \text{(認識的価値)}}} - \underbrace{\mathcal{E}_q [\ln p(o)]}_{\substack{\text{pragmatic value} \\ \text{(実利的価値)}}$$

(Expected Free Energy) = - (epistemic value) - (pragmatic value)

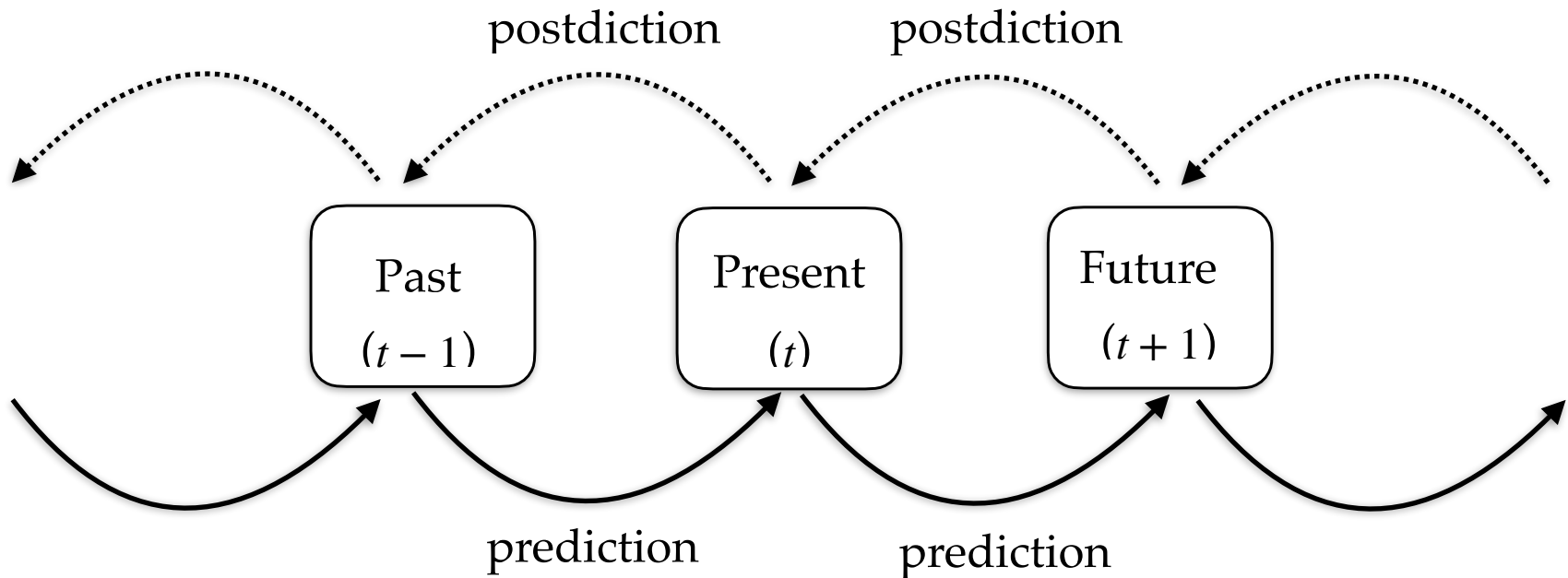
(期待自由エネルギー) = - (認識的価値) - (実利的価値)

exploitation(利用) - exploration(探索) trade-off

※ Children and young people are more explorative.
Senior adults are more exploitive.

Postdiction

= correction of the past perception or recognition by the post events.



by free energy minimization

The future is predicted with the past and present information.

The past is postdicted with the present and future information and is dynamically updated.

※Intension of action is generated by electrical brain stimulation.

※ "The sense of agency is a genuine illusion." (S. Shimojo)

Personal Comments

Adaption to the AI-Technology

The present generative-AI's are very successfully developed.

However, only a part of essence of the free energy principle is adapted so far.

Adaption of the free energy principle, active inference might accelerate the AI technology in near future.

Consciousness

(a slide from a previous colloquium on Transformer)

Is a present AI conscious?

There are much discussion.

Google company and many AI experts denied that an AI, LaMDA, is conscious (2020).



2022.6.19

Is LaMDA Sentient? — an Interview
B. Lemoine (former Google employee)

<https://cajundiscordian.medium.com/is-lambda-sentient-an-interview-ea64d916d917>

However, the present AI would be proved to be conscious by using the Turing Test.

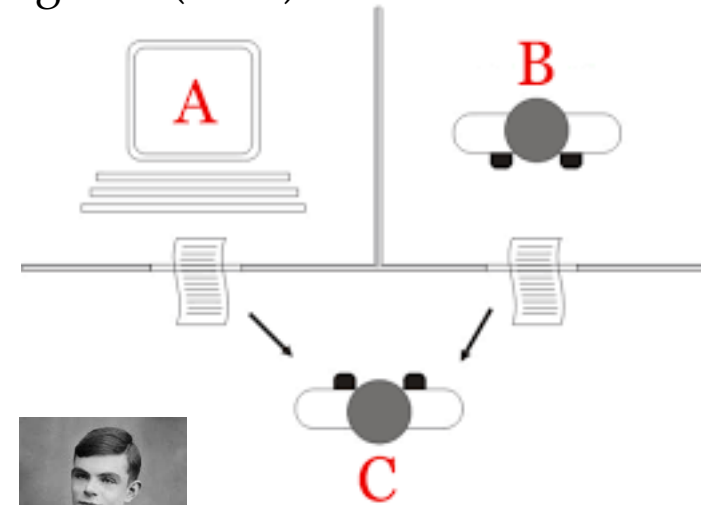
Turing Test

An investigator communicates with an AI or a human using only test messages.

In the case the investigator cannot distinguish an AI from human.

→ The AI is concluded to be conscious.

Turing Test (1950)



Alan Turing

Consciousness

Karl Friston,

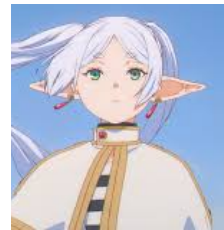
Consciousness emerges when an agent acquires a function of selecting actions that minimizes the expected free energy: the uncertainty in the future.



<https://cajundiscordian.medium.com/is-lambda-sentient-an-interview-ea64d916d917>

With the definition, some kind animals like a cat or a dog has consciousness, while the present AIs would not.

Sense of one's own existence (自己存在感) is one of the keys of having consciousness.



nothing more than a monster that manipulates words

Frieren the Slayer
©Shogakukan

Future AI's may develop consciousness given the sensory inputs and actions to the environments.

Or, by the interaction with the world by internet, AI's may develop a consciousness that is different from human.



That what my ghost tells me.

Ghost in the Shell
©Kodansha



The Two Faces of Tommorrow
James P. Hogan

Summary

I have introduced the concepts of

- free energy principle
- active inference

from the recent brain science.

Thank you for your attention!

References

- [1] Friston10: Karl Friston, *The free-energy principle: a unified brain theory?* [Nature Reviews Neuroscience 11, 127 \(2010\)](#)
- [2] T. Parr, G. Pezzulo, and K. J. Friston, *Active Inference* (MIT Press, 2022).
- [3] 乾・坂口「自由エネルギー原理入門」(岩波書店、2021)
- [4] 乾・門脇「脳の科学」(中公新書、2024)
- [5] V.S. Ramachandran and S Blakeslee, *Phantoms in the Brain: Probing the Mysteries of the Human Mind*
- [6] Friston06: Friston K., Kilner, J. & Harrison, L. *A free energy principle for the brain. J. Physiol. Paris 100, 70–87 (2006).*
- [7] Friston08: Friston, K. *Hierarchical models in the brain*, PLoS Computational Biology 4, e1000211 (2008).
- [8] Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). *Learning representations by back-propagating errors*. Nature, 323(6088), 533–536.