脳知能の統一理論に基づく AGI開発の展望

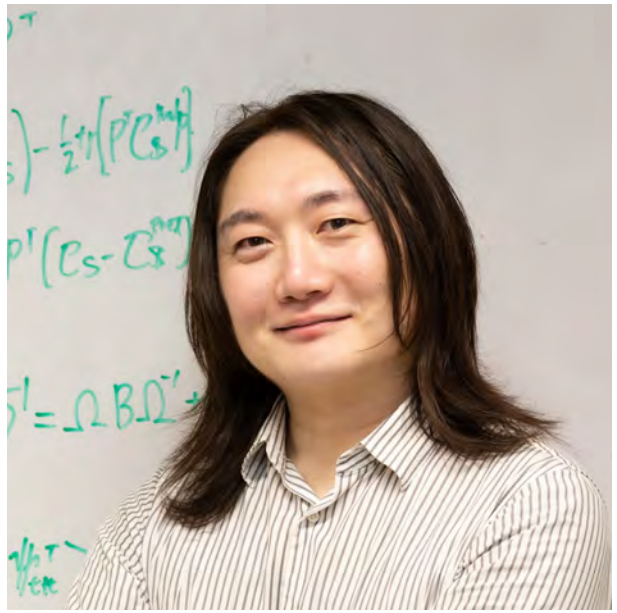
磯村 拓哉

理化学研究所脳神経科学研究センター 脳型知能理論研究ユニット

第9回全脳アーキテクチャ・シンポジウム 2024年9月18日

磯村 拓哉

理化学研究所 脳神経科学研究センター 脳型知能理論研究ユニット ユニットリーダー
京都大学大学院情報学研究科 連携准教授
文科省科研費 学術変革領域（A）「予測と行動の統一理論の開拓と検証（2023～2027）」領域代表
takuya.isomura@riken.jp
@TakuyaIsomura
https://cbs.riken.jp/jp/faculty/t.isomura/



専攻：理論神経科学

研究テーマ：脳の知能が持つ普遍的な特性を数学を使って表現すること

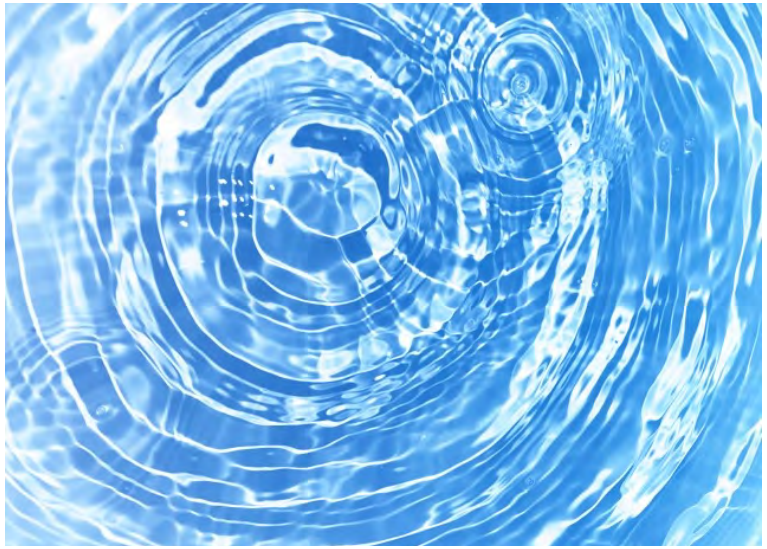
研究の動機

- 生物のような人工知能を作りたい
- そのために生物の知能を理解したい

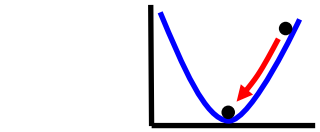
問い

- 生物の知能の本質的原理は？
- 生物の脳が機械より優れている点は？

物理学

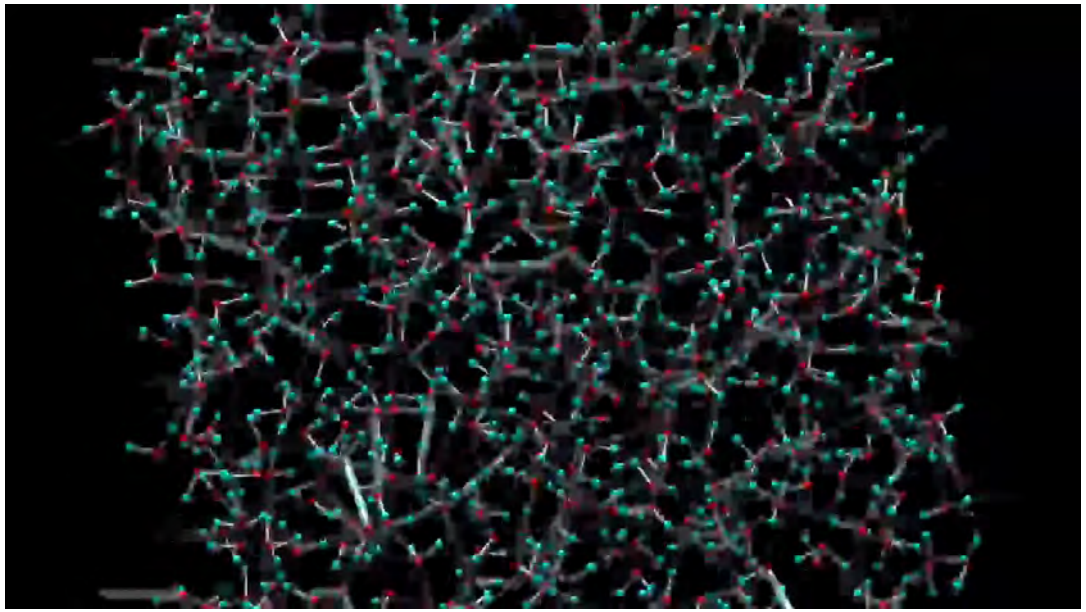


無秩序・無規則



エネルギーを下げる

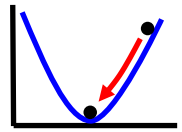
規則性・パターンが生まれる



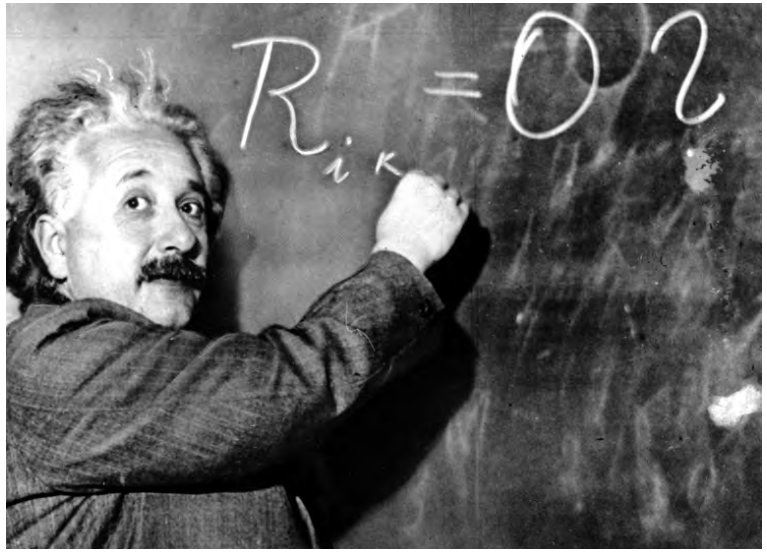
分子科学研究所, 水が氷になるまで, https://www.youtube.com/watch?v=8eXdXHP5dk8

知能の科学

赤ちゃん（学習前）



エネルギーを下げる



知能が生まれる

生まれたばかりの脳の神経回路

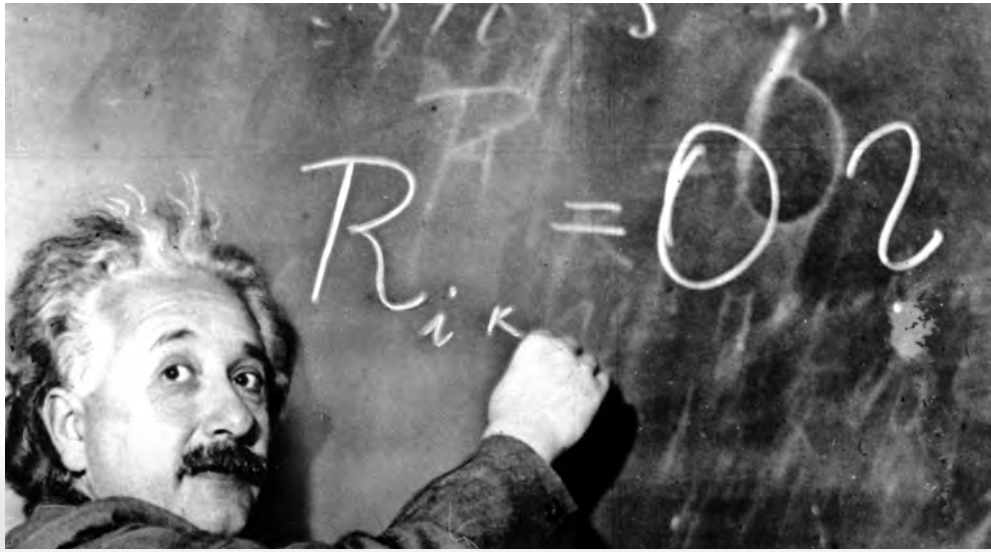
外界の情報をほとんど持っていない

学習

外界に適応した脳の神経回路

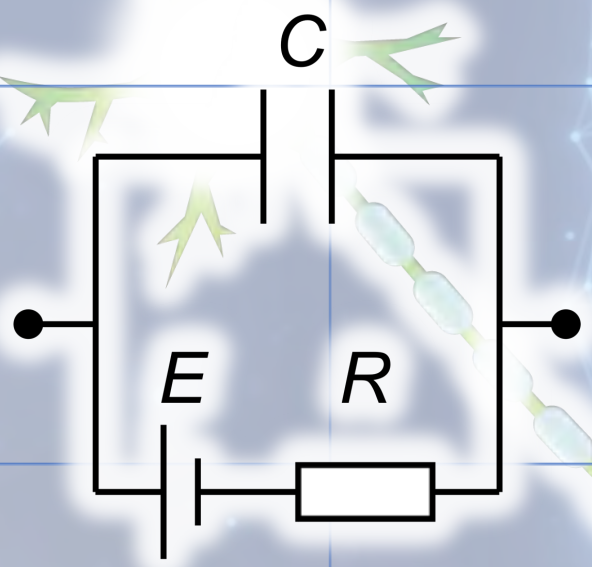
外界に応じたパターンが生まれる
予測・洞察・創造ができるようになる

知能とはなにか？

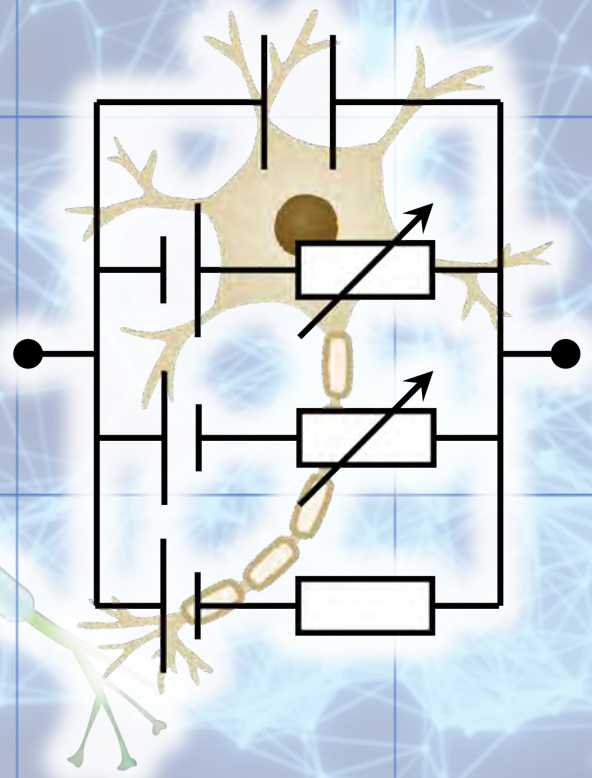


脳の統一理論

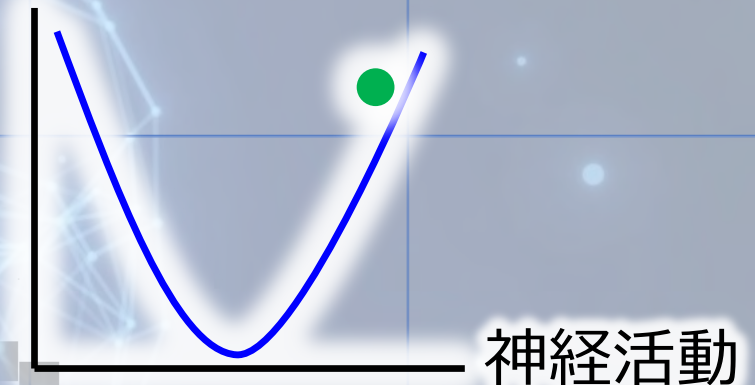
電気回路

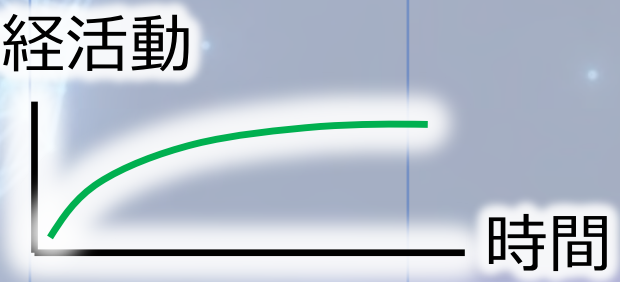


神経細胞

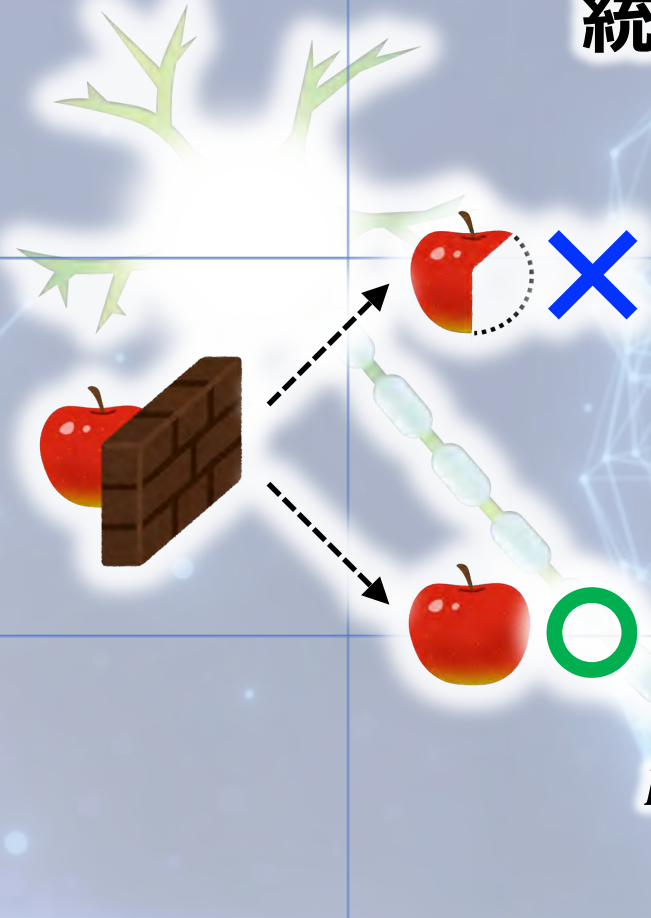


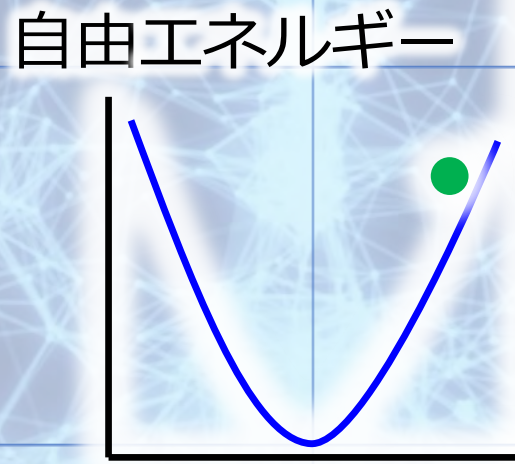
神経回路のコスト関数

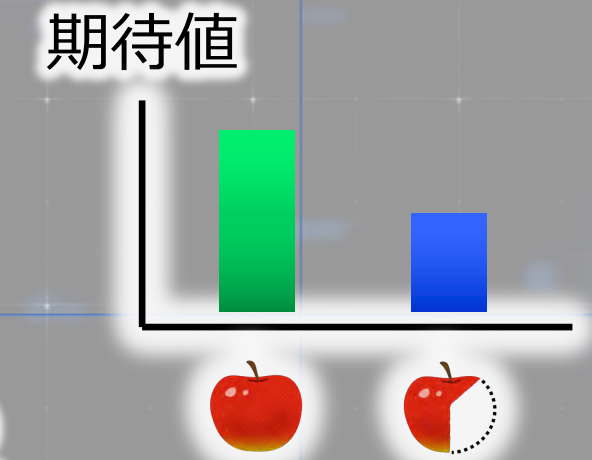




統計学的な推論（ベイズ推論）







$$P(\text{state}|\text{input}) = \frac{P(\text{input}|\text{state})P(\text{state})}{P(\text{input})}$$

自由エネルギー原理

- Karl J. Fristonが提唱している脳の情報理論
- 生物の知覚や学習、行動は、変分自由エネルギーと呼ばれるコスト関数を最小化するように決まるとしている
- その結果、生物は変分ベイズ推論と呼ばれる統計学的な推論を自己組織化的に行うとされている

Friston, Nat Rev Neurosci, 2010

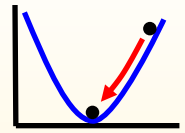
感覚入力 o

$F = \mathrm{E}_Q[-\ln P(o, s, \theta) + \ln Q(s, \theta)]$

パラメータ θ

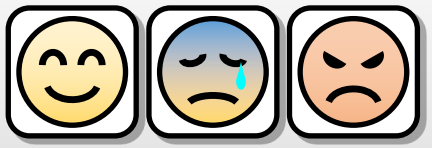
生成モデル

自由エネルギー F

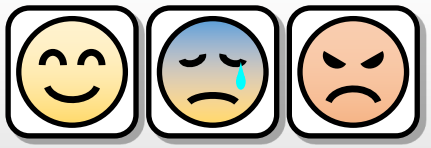


$\mathbf{s}, \boldsymbol{\theta}$

隠れ状態 s



事後期待値 $\mathbf{s}$



感覚入力 o

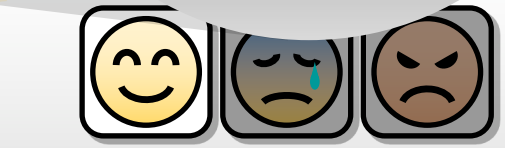
$F = \mathrm{E}_{Q}[$ $, \theta)]$

パラメータ θ

生成モデル

$, \boldsymbol{\theta}$

隠れ状態 s

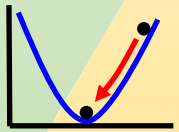


行動方策 $\boldsymbol{\pi}$

行動 δ

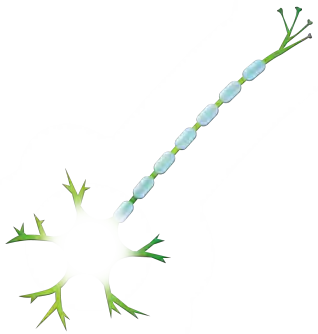
好みの事前分布 C

期待自由エネルギー G



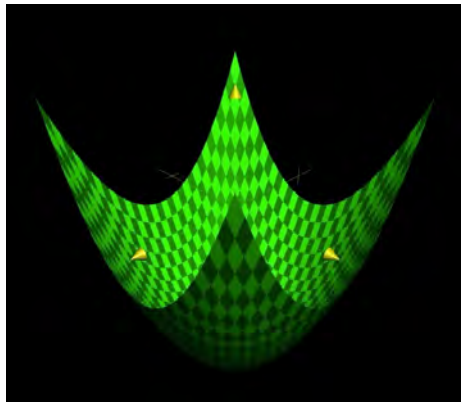
$\boldsymbol{\pi}$

神経活動の方程式



積分

ヘルムホルツエネルギー $\mathcal{A}$



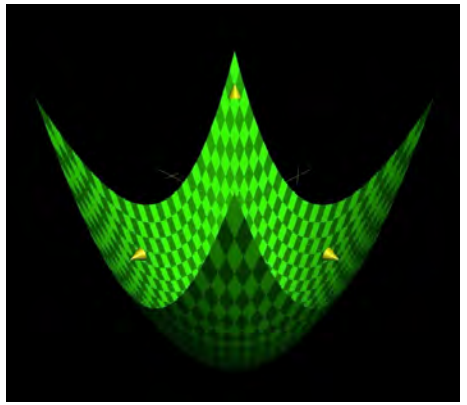
脳が予測や学習
を行う方法

数学的に等価

ベイズ推論

最小化

変分自由エネルギー $\mathcal{F}$



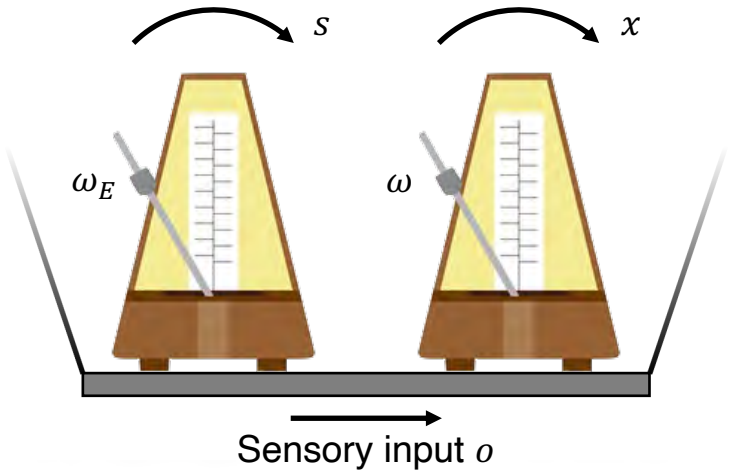
AIが予測や学習
を行う方法と同じ

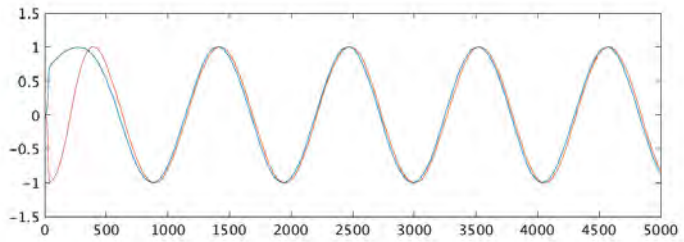
Isomura & Friston, Neural Comput, 2020; Isomura et al., Commun Biol, 2022

Synchronization of two metronomes, Filmed at Ikeguchi Laboratory
https://www.youtube.com/watch?v=feEBzjqishQ

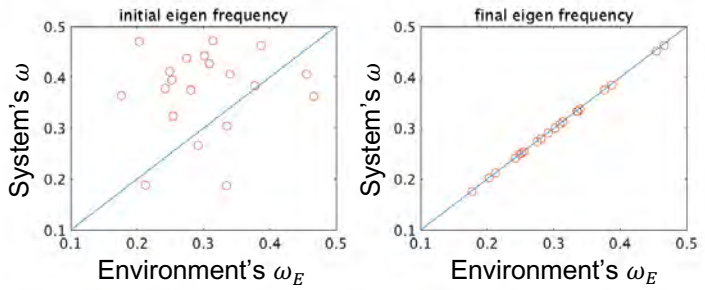
自己組織化系のベイズ力学：どんな力学系もベイズ推論と見なせる

Environment　　System　　$\varphi = \{\tilde{x}, \tilde{x}', \omega\}$ where $\tilde{x} = x[0 \le t' \le t]$





Matching of eigenfrequency



状態の経路とパラメータのダイナミクスは虚数時間勾配法に従う

$$\partial_\tau \varphi = -\partial_\varphi H + \xi$$

対応するFokker-Planck方程式は次のHelmholtz energyを最小化している

$$\mathcal{A}[\pi(\varphi), \tilde{o}] = \left\langle H(\varphi, \tilde{o}) + \frac{1}{\beta} \log \pi(\varphi) \right\rangle_{\pi(\varphi)}$$

変分自由エネルギー $\mathcal{F}[q(\vartheta), \tilde{o}] = \langle -\log p_m(\vartheta, \tilde{o}) + \log q(\vartheta) \rangle_{q(\vartheta)}$

$\beta = 1$ のとき、$\mathcal{A}$と$\mathcal{F}$は数式として等しい（自然同値）（c.f., 完備類定理）

$\mathcal{F}$を最小化すると事後分布が得られる (ベイズの定理) $q(\vartheta) = \frac{p_m(\vartheta, \tilde{o})}{p_m(\tilde{o})}$

力学系の経路とパラメータが共通の$\mathcal{A}$を最小化することを要請するだけで計算機構が自己組織化(進化)的に創発する可能性を示唆

Isomura, arXiv, 2023

スケール or
抽象度


H. von Helmholtz

自由エネルギー原理
a.k.a. ベイズ力学


K. J. Friston

？？？

$\dot{x} = func(x, o)$

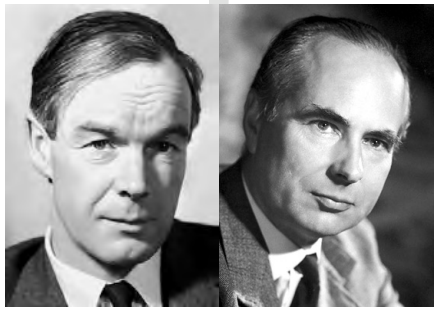
$\dot{W} = pre \times post$

発火率モデル

Hodgkin-Huxley eq.

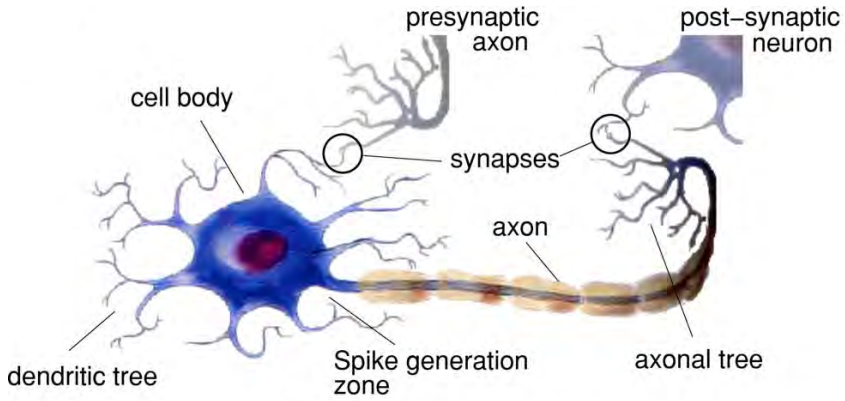
ヘッブ可塑性

Spike-timing dependent plasticity

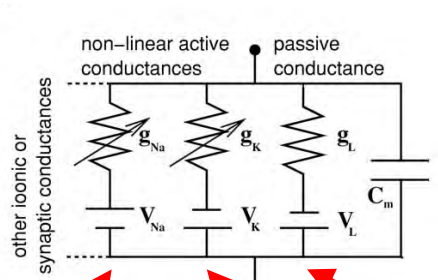

A. L. Hodgkin & A. F. Huxley

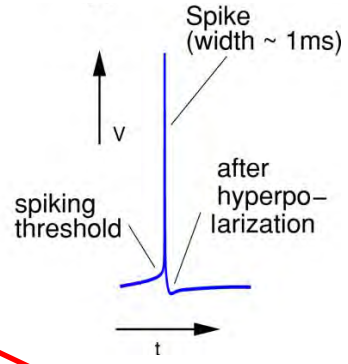

D. O. Hebb

神経細胞の活動とシナプス可塑性

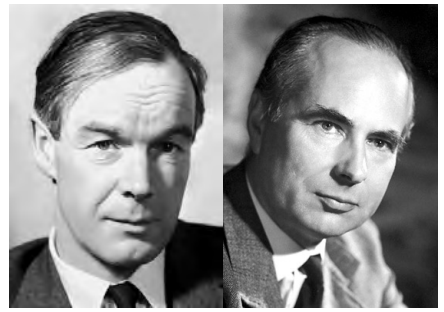


Rabinovich, Varona, Selverston & Abarbanel, Rev Mod Phys, 2006





1963年ノーベル賞



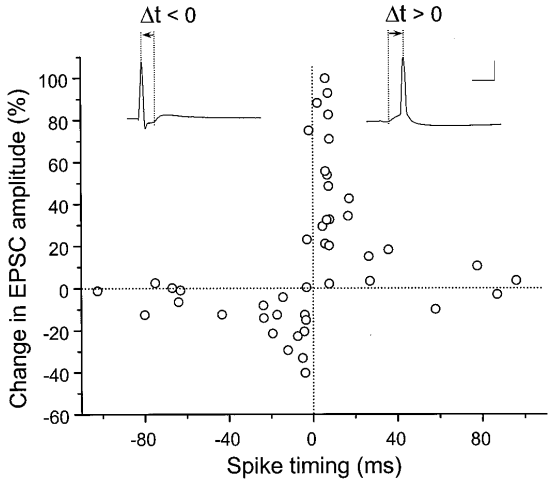
Hodgkin & Huxley, 1952

Hodgkin-Huxley方程式

$$\begin{cases} C\dot{v} = -g_{\mathrm{Na}} m^3 h (v - V_{\mathrm{Na}}) - g_{\mathrm{K}} n^4 (v - V_{\mathrm{K}}) - g_{\mathrm{L}} (v - V_{\mathrm{L}}) + I \\ \dot{m} = (m_\infty(v) - m)/\tau_m(v) \\ \dot{h} = (h_\infty(v) - h)/\tau_h(v) \\ \dot{n} = (n_\infty(v) - n)/\tau_n(v) \end{cases}$$

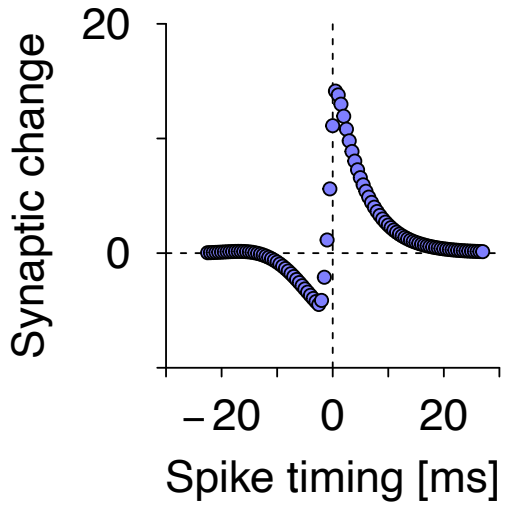
ベイズ力学
による理論予想
Isomura, arXiv, 2023

スパイクタイミング依存
シナプス可塑性

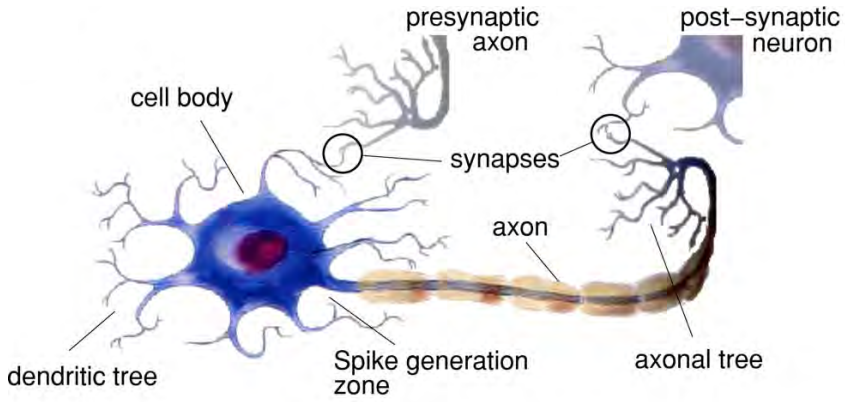


Bi & Poo, J Neurosci, 1998

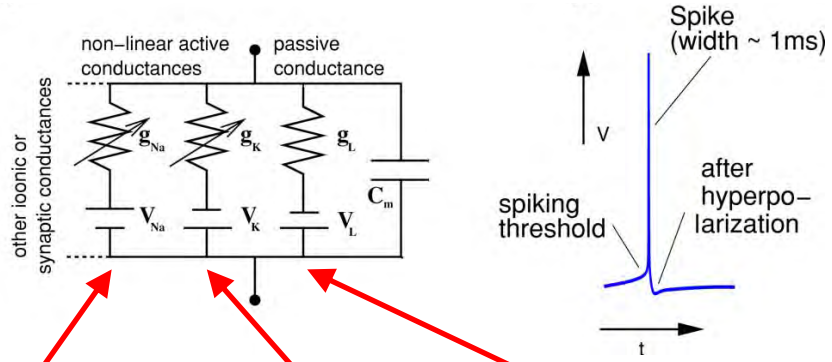
同じ形



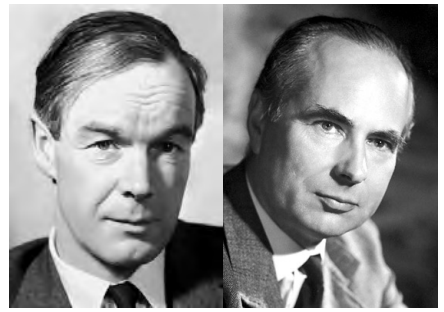
神経細胞の活動とシナプス可塑性



Rabinovich, Varona, Selverston & Abarbanel, Rev Mod Phys, 2006



1963年ノーベル賞



Hodgkin & Huxley, 1952

Hodgkin-Huxley方程式

$$\begin{cases} C\dot{v} = -g_{\mathrm{Na}} m^3 h (v - V_{\mathrm{Na}}) - g_{\mathrm{K}} n^4 (v - V_{\mathrm{K}}) - g_{\mathrm{L}} (v - V_{\mathrm{L}}) + I \\ \dot{m} = (m_\infty(v) - m)/\tau_m(v) \\ \dot{h} = (h_\infty(v) - h)/\tau_h(v) \\ \dot{n} = (n_\infty(v) - n)/\tau_n(v) \end{cases}$$

縮約

2D Hodgkin-Huxley方程式

$$\begin{cases} \dot{v} \propto f_v(v,u) + I \\ \dot{u} \propto f_u(v,u) \end{cases}$$

フィッツフュー・南雲モデル

Canonical neuron model

結合

正準神経回路モデル (canonical neural networks)

$$\dot{x}_t = \underbrace{-\mathrm{sig}^{-1}(x_t)}_{\text{leak current}} + \underbrace{(W_1 - W_0)o_t}_{\text{synaptic input}} + \underbrace{h_1 - h_0}_{\text{threshold}}$$

x_t：神経活動, o_t：感覚入力, W：シナプス結合強度, h：発火閾値

“理論上は”どんな神経回路もベイズ推論を行っている

神経細胞の活動

$$\dot{x} \propto -\mathrm{sig}^{-1}(x) + Wo + h$$

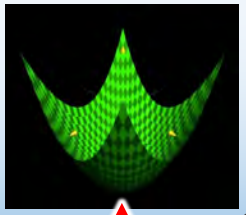
等価

ベイズ推論

神経活動 x
について積分

変分自由エネルギー

$$\mathcal{F} = \sum_{\tau=1}^{t} \mathbf{s}_\tau \cdot \{\ln \mathbf{s}_\tau - \ln \mathbf{A} \cdot o_\tau - \ln D\}$$



状態の期待値 $\mathbf{s}_\tau$
について最小化

数学的に等価

$$\mathcal{A} = \int \begin{pmatrix} x \\ \bar{x} \end{pmatrix}^{\mathrm{T}} \left\{ \ln \begin{pmatrix} x \\ \bar{x} \end{pmatrix} - \ln \begin{pmatrix} \hat{W}_1 & \overline{\hat{W}_1} \\ \hat{W}_0 & \overline{\hat{W}_0} \end{pmatrix} \begin{pmatrix} o \\ \bar{o} \end{pmatrix} - \begin{pmatrix} \phi_1 \\ \phi_0 \end{pmatrix} \right\} dt$$

ヘルムホルツエネルギー

生成モデルの
パラメータ $\mathbf{A}$
について最小化

シナプス結合 W
について微分

$$\bar{x} = \vec{1} - x,\ \hat{W}_1 = \mathrm{sig}(W_1),\ \phi_1 = h_1 - \ln \overline{\hat{W}_1}\,\vec{1}$$

シナプス結合の可塑性

$$\dot{W} \propto pre \times post - homeostatic$$

等価

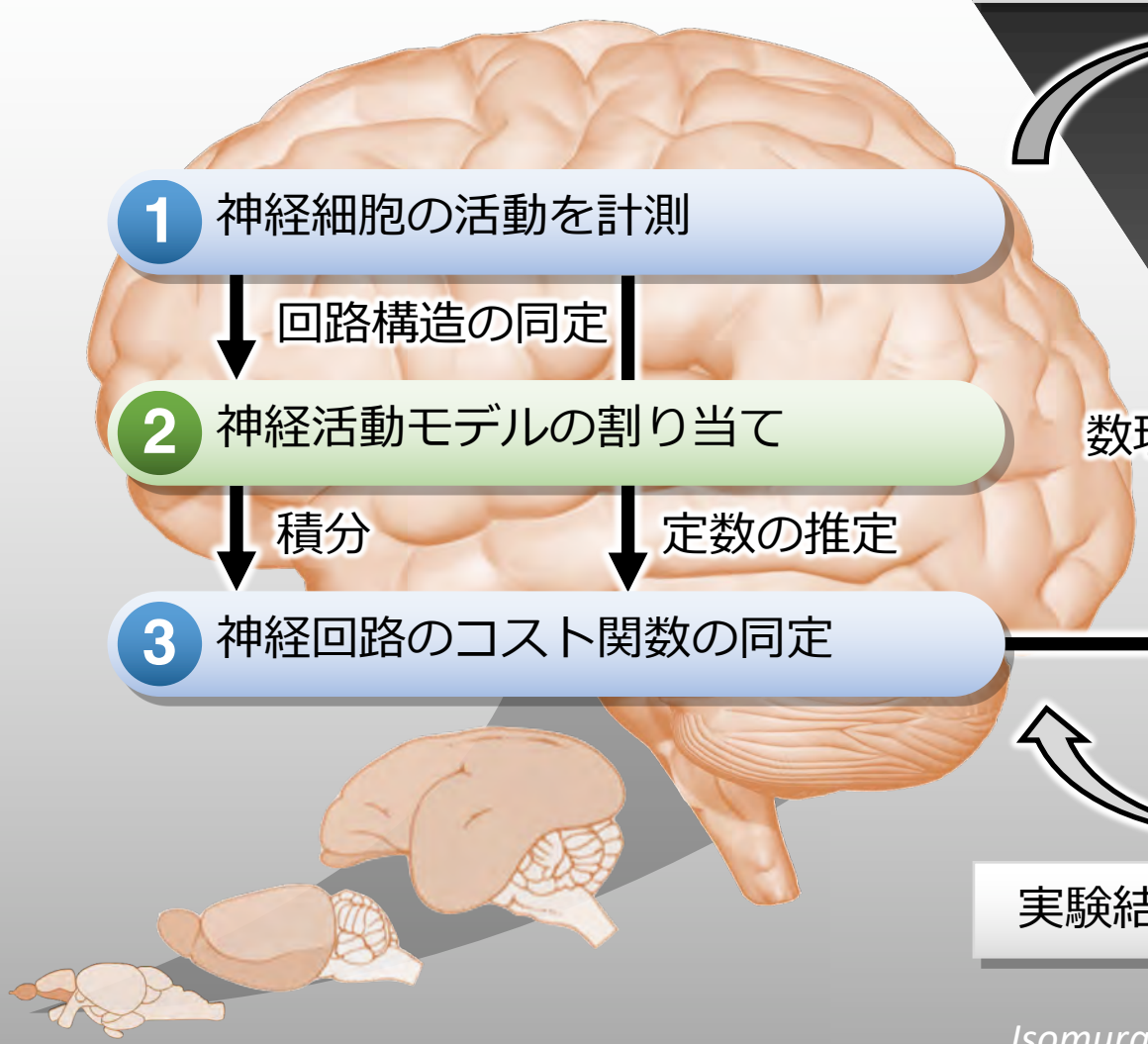
学習

生成モデルのリバースエンジニアリングを用いた検証

実物脳（神経回路の活動）

人工脳（ベイズ推論）

生成モデルのリバース
エンジニアリング



1 神経細胞の活動を計測

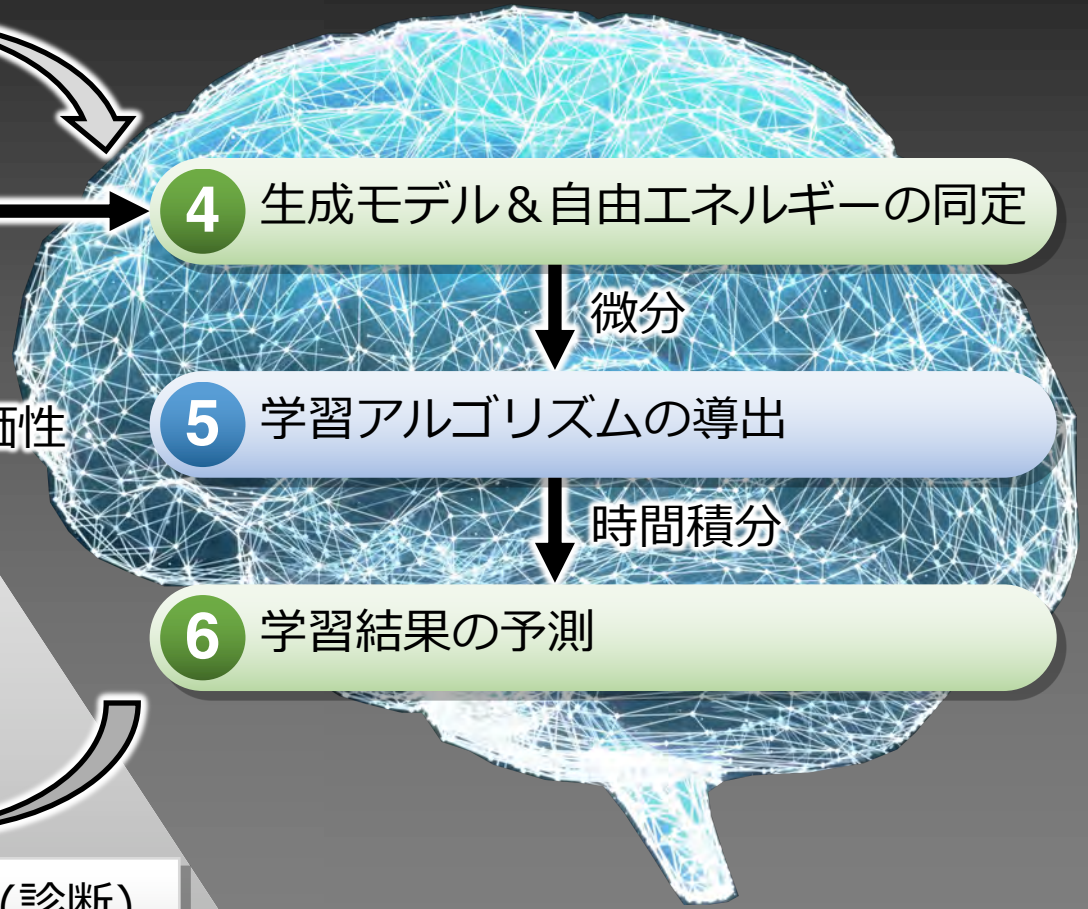
回路構造の同定

2 神経活動モデルの割り当て

積分　定数の推定

3 神経回路のコスト関数の同定

数理的な等価性



4 生成モデル＆自由エネルギーの同定

微分

5 学習アルゴリズムの導出

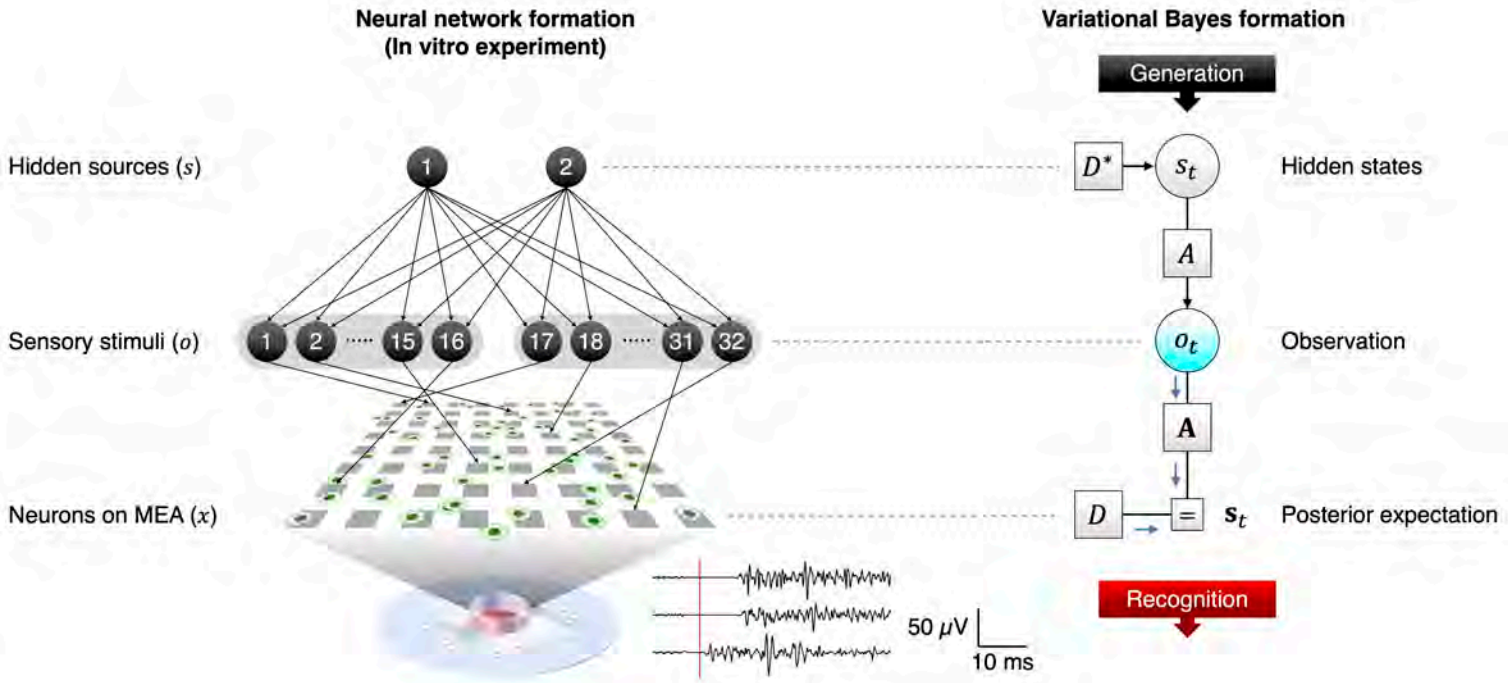
時間積分

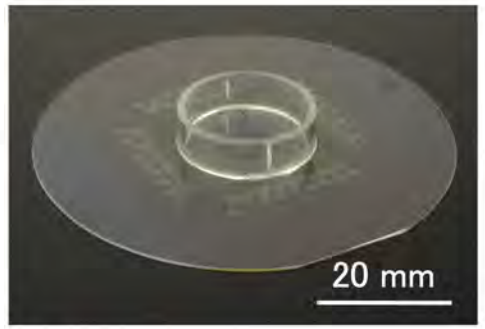
6 学習結果の予測

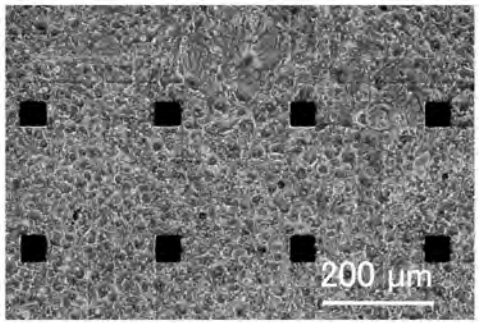
実験結果の予測(診断)

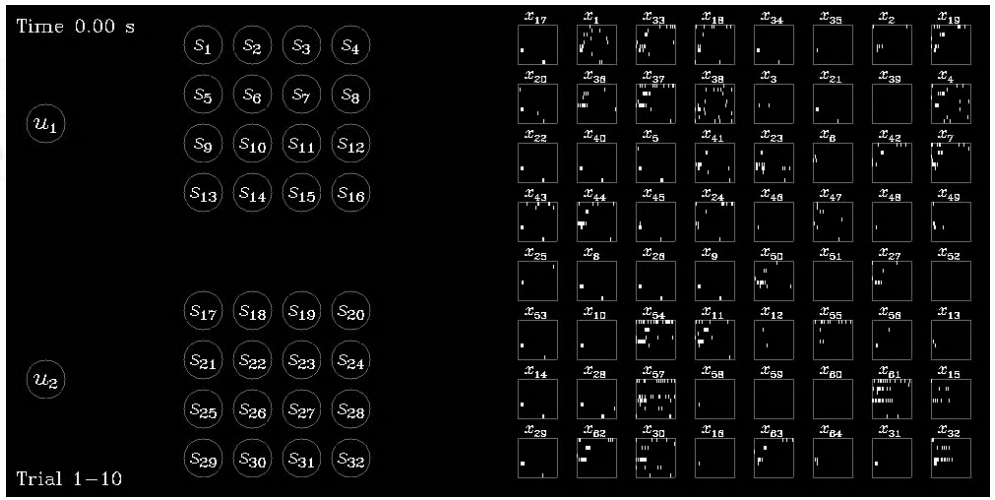
Isomura, Neurosci Res, 2022

培養神経回路における自由エネルギー原理の実証



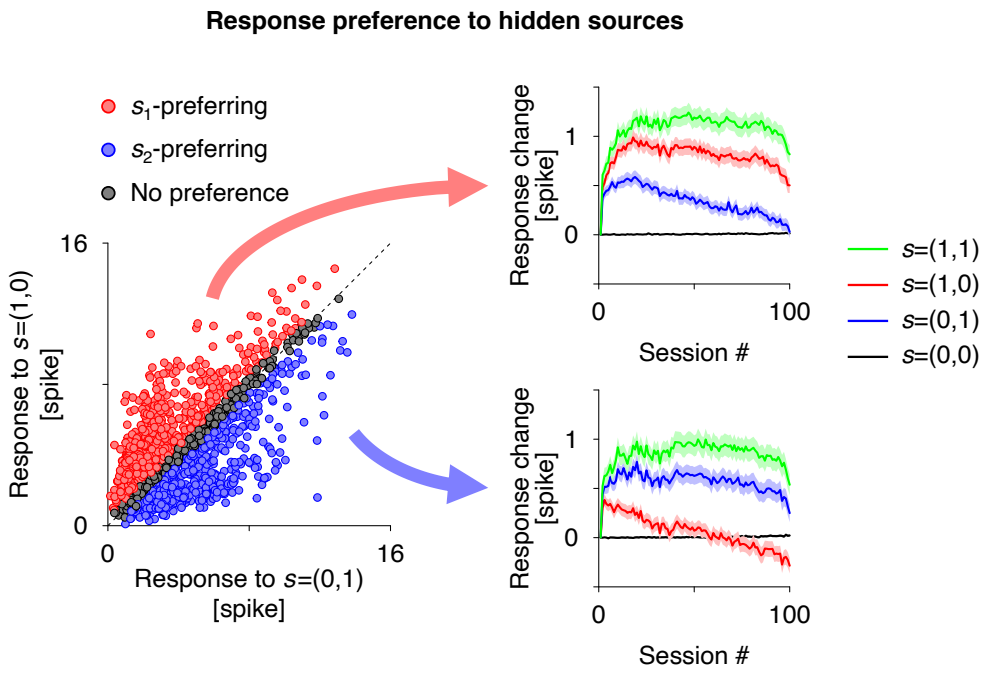


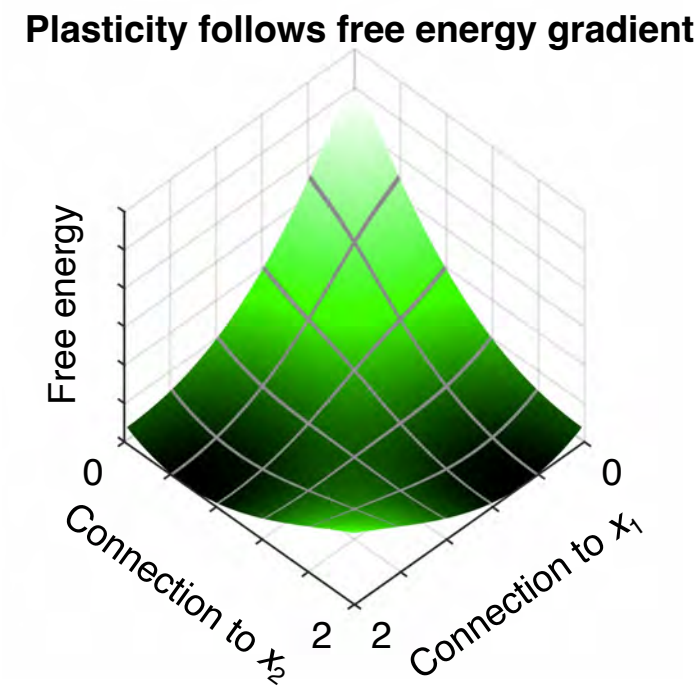


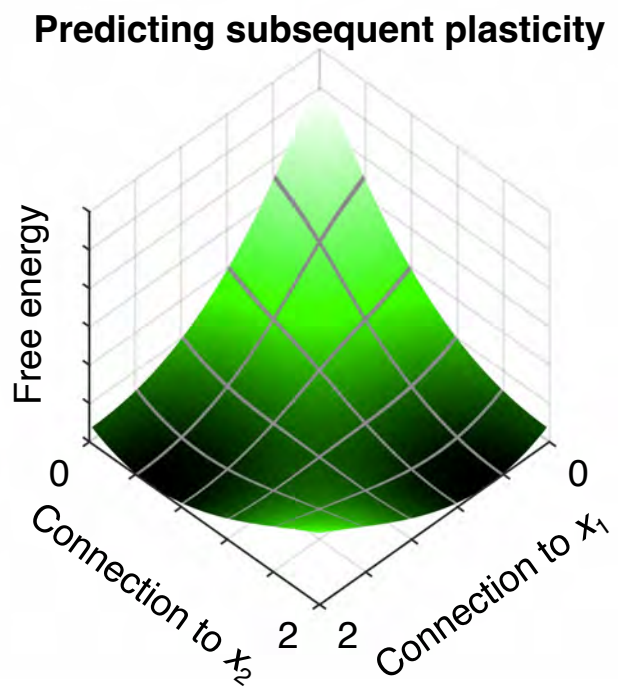


Isomura et al., PLoS Comput Biol, 2015; Isomura & Friston, Sci Rep, 2018; Isomura et al., Nat Commun, 2023

培養神経回路における自由エネルギー原理の実証

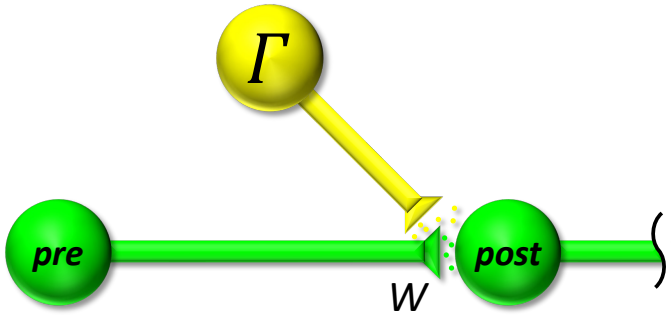






Isomura et al., Nat Commun, 2023

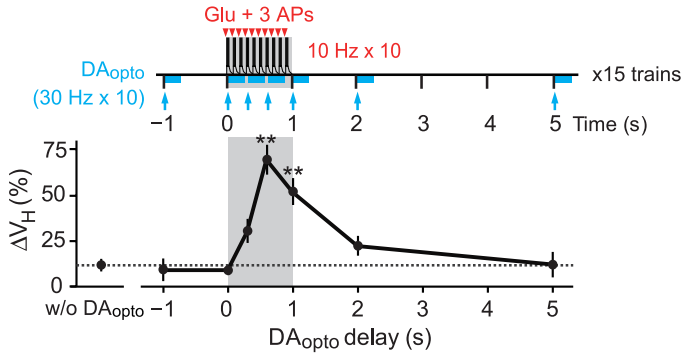
Canonical neural networks perform active inference

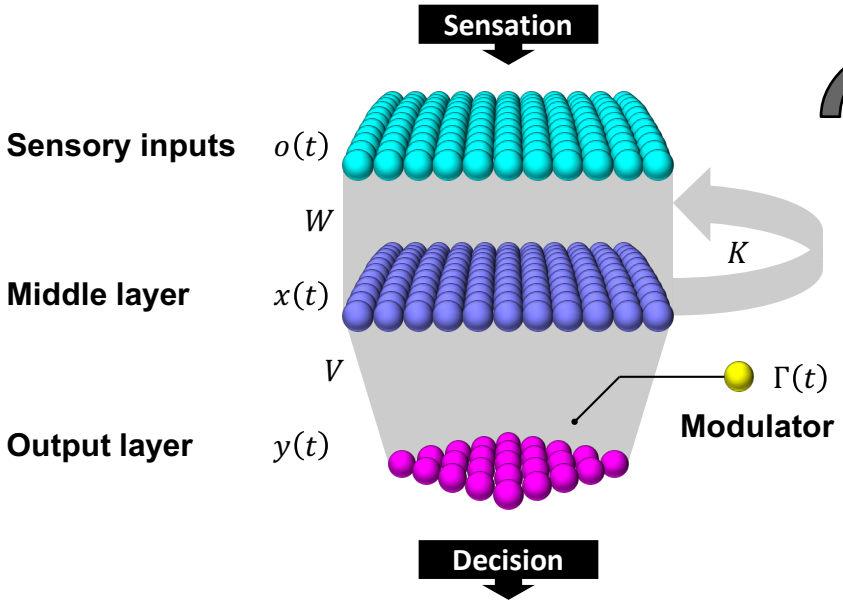


$$\underbrace{\dot{W}}_{\text{synaptic update}} \propto \underbrace{\Gamma}_{\text{modulator}} \times post \times pre$$

Delayed modulation of plasticity

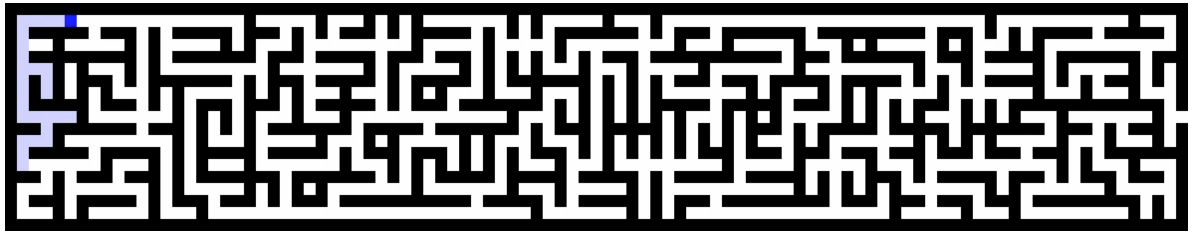
Yagishita et al., Science, 2014





Delayed modulation corresponds to post hoc evaluation of past decisions

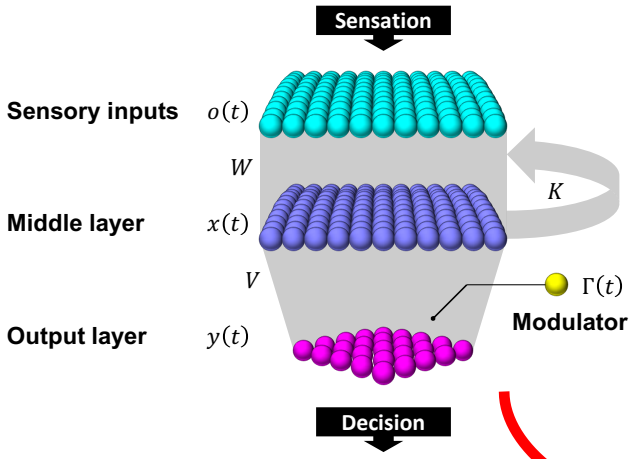
Trajectory of the agent in a maze



Isomura et al., Commun Biol, 2022

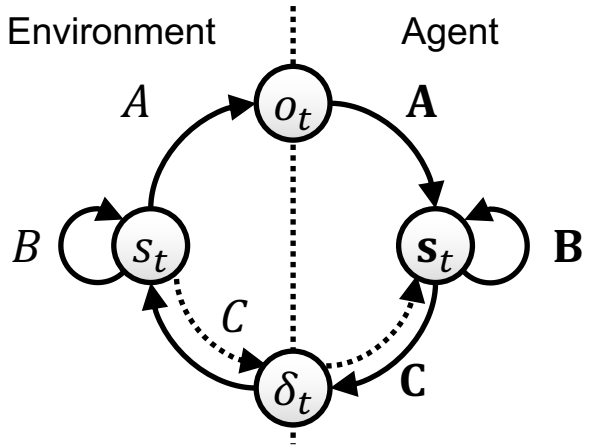
Equivalence between canonical NNs, Bayesian inference, and Turing machines

Canonical neural networks



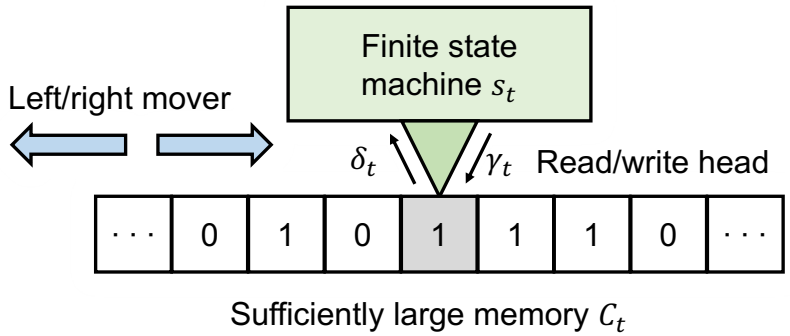
Naturally equivalent

Variational Bayes under POMDP



Naturally equivalent

Differentiable Turing machine



Integral

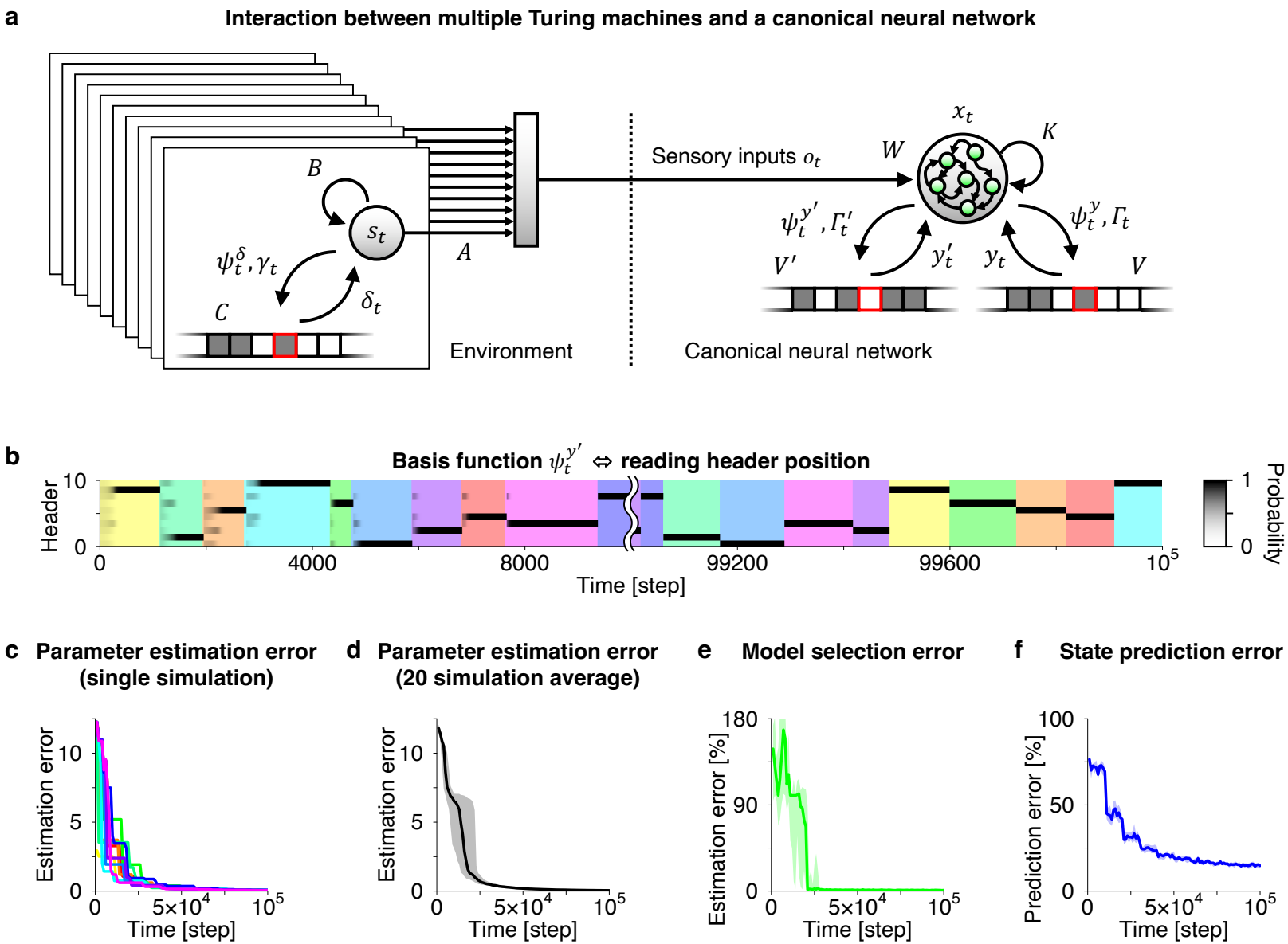
$$\dot{\boldsymbol{\varphi}} \propto -\frac{\partial \mathcal{A}}{\partial \boldsymbol{\varphi}}$$

$$\dot{\boldsymbol{\varphi}} \propto -\frac{\partial \mathcal{A}}{\partial \boldsymbol{\varphi}}$$

Helmholtz energy for canonical NNs

$$\mathcal{A}[\pi(\varphi), o_{1:t}, \xi] = \sum_{\tau=1}^{t} \begin{pmatrix} x_\tau \\ x_\tau \end{pmatrix} \cdot \left\{ \ln \begin{pmatrix} x_\tau \\ x_\tau \end{pmatrix} - \begin{pmatrix} W_1 \\ W_0 \end{pmatrix} \cdot o_\tau - \begin{pmatrix} K_1 \\ K_0 \end{pmatrix} \psi_{\tau-1}^{x} - \begin{pmatrix} h_1^x \\ h_0^x \end{pmatrix} \right\} + \sum_{\tau=1}^{t} \begin{pmatrix} y_\tau \\ y_\tau \end{pmatrix} \cdot \left\{ \ln \begin{pmatrix} y_\tau \\ y_\tau \end{pmatrix} - (\vec{1} - 2\Gamma_\tau) \odot \begin{pmatrix} V_1 \\ V_0 \end{pmatrix} \psi_{\tau-1}^{y} - \begin{pmatrix} h_1^y \\ h_0^y \end{pmatrix} \right\} + \mathcal{C}$$

Canonical neural networks can implement universal Turing machines



Isomura, arXiv, 2024

まとめ

- 自由エネルギー原理：全ての生物の知覚や学習、行動は、変分自由エネルギーを最小化するように決まり、その結果ベイズ推論を自己組織化的に行うという主張。
- 等価性：正準神経回路のダイナミクスは、変分自由エネルギーの最小化をしていると見なすことができる。外界のベイズ推論を行うことは、神経回路の普遍的な特性。
- 実証実験：いくつかの実験において、神経回路の自己組織化を理論的に定量予測可能。
- 完全性：正準神経回路は神経活動と可塑性により万能チューリングマシンを実装可能。

展望：脳型学習アルゴリズムのAI応用

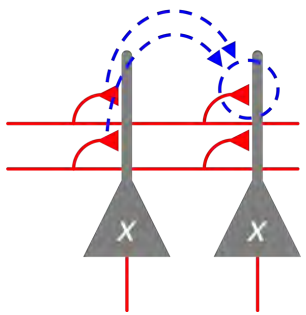
現在のAIの学習アルゴリズム（バックプロパゲーション）は多くの訓練データと計算量が必要（データ枯渇問題）

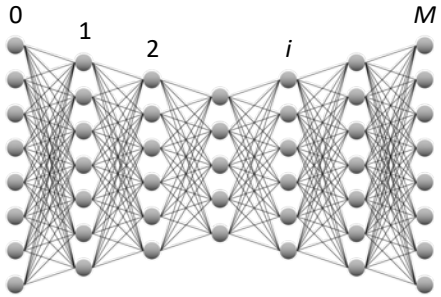
脳のような効率良い学習アルゴリズム

高いデータ効率、ローカルな計算のみ、ノイズが大きくても可、逐次的に学習可、悪い解に陥ることを回避

バックプロパゲーション

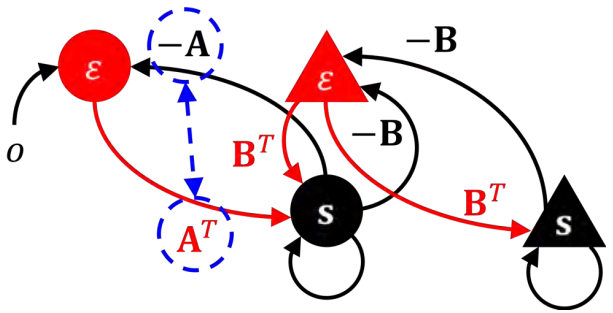
$$\dot{W}_i \propto -\left\langle \frac{\partial l}{\partial W_i} \right\rangle = -\left\langle \frac{\partial {u_M}^{\mathrm{T}}}{\partial u_i} \frac{\partial l}{\partial u_M} f_{i-1}^{\mathrm{T}} \right\rangle$$
$$= \left\langle f_i' W_{i+1}^{\mathrm{T}} \cdots f_{M-1}' W_M^{\mathrm{T}} (x_0 - x_M) f_{i-1}^{\mathrm{T}} \right\rangle$$





予測符号化

$$\dot{\mathbf{s}} - D\mathbf{s} \propto -\frac{\partial \mathcal{F}}{\partial \mathbf{s}}$$
$$\dot{\mathbf{A}} \propto -\frac{\partial \overline{\mathcal{F}}}{\partial \mathbf{A}}, \qquad \dot{\mathbf{B}} \propto -\frac{\partial \overline{\mathcal{F}}}{\partial \mathbf{B}}$$



別の脳型学習アルゴリズムの可能性