End-to-End 強化学習による知能創発と「思考」創発へ向けた新しい強化学習
Emergence of Intelligence through End-to-End Reinforcement Learning and New Reinforcement Learning toward the Emergence of "Thinking"

大分大学　○ 柴田 克成
Katsunari Shibata
Oita University

Abstract　On the ground of our 20-year relevant researches, I have propounded the significance of "End-to-End Reinforcement Learning" using a recurrent neural network from the viewpoint of emergence of intelligence. Toward the ambitious goal of emergence of "Thinking", I will also introduce my new fundamental idea that "Explorations" and "Actions" should be unified and the internal "Exploration" grows into "Thinking" through learning. New reinforcement learning I have proposed to realize the idea uses a chaotic neural network and jumps out of the conventional stochastic action selection approach.

1　脳の理解をあきらめろ !?

近年，深層学習 (Deep Learning) を中心とした人工知能 (AI) が大流行りである。事の本質は一体どこにあるのであろうか？多くの人が畳み込み (Convolution) 構造を挙げるであろう。しかし，20 年以上この手の研究をしてきた者としては [1][2]，やはり一番大きなポイントは「脳の処理の理解をあきらめ，学習可能な並列マシンに任せた方がいい」ということではないかと考えている。

脳は千数百億個ものニューロンが，平均数万個ものシナプスを介して直接情報のやり取りをし，並列処理を行なっていると言われる [3]。さらに，ニューロンは成長し [4]，柔軟にその機能を分担し合い [5][6]，かつ全体の調和がとれている恐るべきシステムである。

一方，われわれが脳の処理を理解するためには，それは無意識ではなく，当然意識に上る必要がある。しかし，「妻と義母」などの多義図形の認識からもわかるように，通常われわれは一度に一つのことしか意識に上らない。この際，脳内ダイナミクスは一つのアトラクタに引き込まれることで，他の場合と区別ができるようになる。それが言語などと結び付いてシリアルで逐次的な形で表象され，意識に上るのだと筆者は考えている。このようなシリアルな「意識」の上で超並列な脳の処理を理解しようとしても，それは極めて困難と言わざるを得ない。

その証拠に，普通の人は自分の脳を使って犬と猫の画像をいとも簡単に区別できるが，どんなに優秀な人でも，どうやってその区別をしているかをきちんと説明できる人はいない。つまり，図 1 のように，ほとんど意識に上らない脳での超並列処理自体は，それを意識によって理解したものとは全く違うものとなる。たとえばそのギャップによって錯視が起こるが，そもそも自分の理解が実際の処理と全く違うということ自体に気が付かないため，両者のずれを不思議だと感じてしまうのである。

このように考えると，「人間の知性」の解明をしようと思ったら，「人間による理解」を排除して，並列処理の学習に任せ，どういう仕組みで学習していくかという別の観点からの理解を試みる必要があると考えられる。

従来の画像認識の分野では，明るさの調整をしたり，認識対象を切り出したりと様々な処理をした後，識別の手がかりになる「特徴」を与え，最後に識別を行なってきた。しかし良く考えれば，対象を切り出すためにはある程度その対象の認識が必要であるし，人間が与えた「特徴」に限定して識別する必要もない。

これに対し，深層学習による画像認識では，犬と猫の区別も説明できない人間による「おせっかい」を排除し，並列処理の学習が可能で，脳の神経回路網をお手本としたニューラルネットの学習に全てを任せたのである。その結果，人間が手で設計したものより人間に近い認識ができるようになったということは，まさに「人間による理解の排除」の有効性を裏付けていると言える。

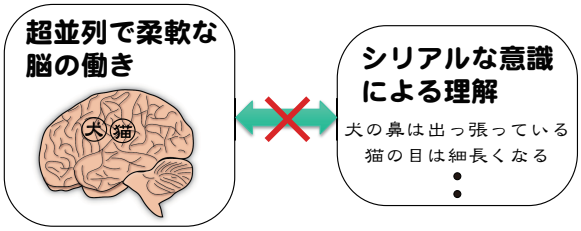

図 1: 全く違う超並列で無意識下の脳の働きとシリアルな意識によるその理解（犬と猫の識別方法の例）

2 “What” 問題から逃げてはいけない

論文 [7] での深層学習による画像認識の例では，夜の街並みや車を背景とした男女が乗ったスクーターの画像を千個のカテゴリーの中から “motor scooter” と正しく識別している。人間にとっては簡単な認識であるが，これだけ複雑な画像を認識できるようになったことは感慨深い。しかしその一方で,「その画像を見て，motor scooter としか認識しないの？」と逆にツッコミを入れたくなる。

このことは，“How” 問題と “What” 問題として捉え直すことができる。たとえば図 2 のように，野球で打球をキャッチするという「予測」が必要な問題を考えよう。従来のアプローチでは,「どうやって打球が落ちる場所を予測するか？」を問題とする。つまり，“How” が解決 (学習) の対象である。このとき，そもそも「何を予測すべきか？」（ここでは打球の落下点）や，前段落での「何を認識すべきか？」というより根源的で知性を必要とする “What” 問題の答えは予め人間が与えてしまい，そこに正面切って取り組むことから逃げてきたのである。実はこれも，元を正すと「人間の理解」に行き着く。

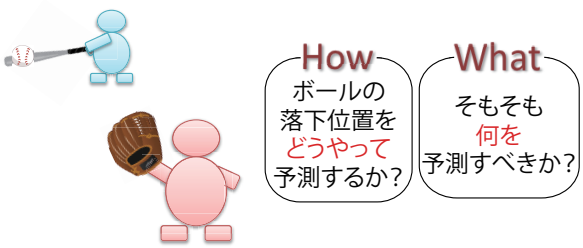

図 2: How 問題と What 問題

3 もっと人間の手を排除すべし

われわれはシリアルな意識で脳の働きを理解するため，脳の処理に「認識」「動作計画」「制御」などの「機能」というラベルを付ける。そして,「認識」して「動作計画」し，最後に「制御」するというように手順を明確にすることで脳の処理を分析し，理解したつもりになる。脳の領野間も非常に密に結合されているにも関わらず，この「機能」をモジュール，つまり，他と分離した個別の処理と捉え，それを統合することで脳の説明をしたり，人間のような知能を形成できると錯覚してしまう。このことは，前述の画像認識過程の細分化と良く似ている。

個々の処理に分割すると，今度はその入出力，つまり “What” 問題の解を予め規定する必要が生じる。しかし，それを決めるためには，結局，処理全体の理解が必要になるというニワトリと卵の関係に陥ってしまう。そして

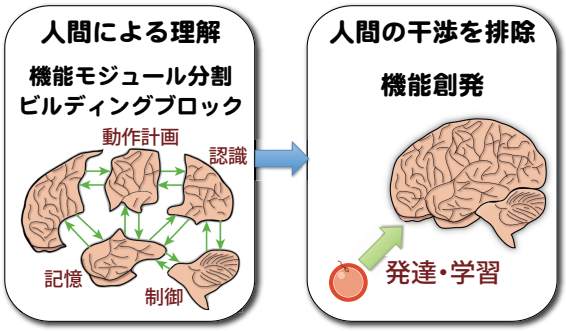

図 3: 「人間の知性」解明に向けた機能創発アプローチへの転換

そのジレンマから抜け出すために，人間がシリアルな意識の上で個々の機能の入出力を規定できるレベルになるまで解くべき問題の範囲を大幅に限定する。こうしてシステムの自由度は大きく拘束され，フレーム問題 [8] を起こし,「高次機能」や「汎用人工知能」(AGI, Artificial General Intelligence)[9] の研究を阻害して来たのである。

そこで筆者らは，図 3 のように,「人間の知性」の解明のためには，機能モジュール分割に基づいた脳の働きの「人間による理解」を排除し，機能がいかに創発するかという発達・学習の観点からの解明が重要と考え，主に学習の観点から研究を進めてきた。そして，深層学習が画像認識で細分化された処理の垣根を取り払ったのと同様に，図 4 のように，センサからモータまでの全ての処理をニューラルネットで構成し，それを強化学習によって試行錯誤 (探索) をもとに自律的に学習させ，内部に機能を創発させる枠組みを提唱して来た [2][10]。

その際,「機能」というものの捉え方を大きく転換した。機能とは「予め定義して与えるもの」ではなく，学習によって一貫性を持って最適化された処理全体に対し,「人間の理解のために後付けした単なるラベル」と考

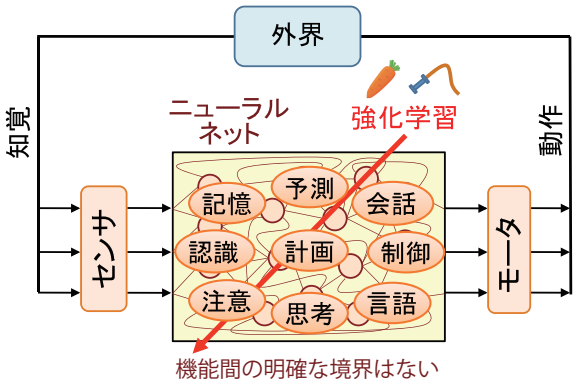

図 4: End-to-End 強化学習による機能創発

えるのである。こうすることで，”What 問題”に対しても予め答えを与える必要がなくなり，一貫性を持った処理全体に対する学習によって解決されることが期待できる。このアプローチを，近年は，DeepMind の Hassabis に倣って「End-to-End 強化学習」[11] と呼んでいる。

4　ものぐさ手法で知能が創発 !?

End-to-End 強化学習は，報酬や罰という最小限の情報を与えるだけで，人間ができるだけ手を出さずに，学習者の試行錯誤に基づく自律学習に処理全体の獲得を任せるという究極なものぐさ手法である。報酬や罰と人間のような知能とのあまりにも大きなギャップのため，たったこれだけで本当に様々な機能が創発するのか？という疑問が生ずる。これに２０年以上掛けてチャレンジして来たのが，筆者らが行って来た研究の大半である。スペースの関係で，詳細は [10]，および，そこに挙げた個別の参考文献を参照していただくとして，ここでは，なぜ機能が創発するのか，そして，具体的にどのような機能が創発したかをざっと紹介していく。

例として，AIBO ロボットが相手の AIBO とキスすると報酬がもらえるタスクを学習した場合 [12] を考える。この場合，センサ信号として画素数を落とした AIBO のカメラ画像をそのままニューラルネットに入力し，その状態と行動に対する評価 (行動価値) を出力する。つまりニューラルネットは，重み値という可変パラメータを有し，カメラの画像から行動価値を出力する関数である。そして，その出力を元に試行錯誤として確率的な行動選択をし，相手の AIBO とキスをすると報酬が与えられる。Actor-Critic または Q-learning という強化学習の手法 [13] は，報酬や罰をもらえない状態においても，その状態がどれぐらい良いかの評価とその評価を上げるための行動を学習する。これに基づいて，毎ステップ教師信号を自動生成し，誤差逆伝播 (BP, BackPropagation) 法によってニューラルネット全体を 1 回だけ学習する。

この学習によって，報酬や罰を元に，適切な評価と行動をするために並列処理を行うニューラルネットの内部は最適化され，役割分担が進む。相手の AIBO が目の前にいれば評価は高く，遠くにいれば低くなっていく。また，左側に相手がいれば左，右側にいれば右を向こうとし，正面にいれば前進しようとする。このとき，周りに AIBO 以外のいろいろなものが置いてあれば，適切な評価と動作を出力するには，画像から AIBO を認識し，どこにいるのかを判断する必要が生ずる。

しかしこのとき，まず認識ありきではなく，認識出力が何かも規定しない。画像から評価や行動を出力する関数としてのニューラルネットがあり，その内部パラメータとしての重み値を適切な評価や行動を出力するように最適化した結果，認識という機能が創発したと後から解釈を付すのである (図 5)。認識以外の機能も同様であり，この機能はこの部分と明確に特定することはできない。

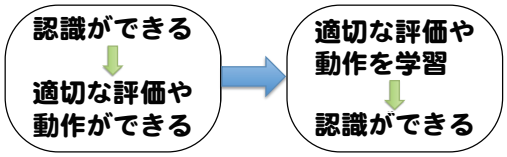

図 5: 「機能」に関する逆転の発想の必要性

この AIBO の学習は，筆者が知る限り，カメラ画像に様々なものが映る中で，画像認識や行動計画の方法を一切与えずに実ロボットに目的達成動作を学習させた世界で初めての例である。またそれ以外にも，階層型ニューラルネットを用いた学習で，センサ動作，Hand-eye Coordination，コミュニケーションなどの機能が，簡単なタスクではあるが，直接的な方法や教師信号等を外部から与えることなく，強化学習によってニューラルネット内部に創発した。また，色の恒常性の錯視や知識転移なども，動作信号の学習からある程度説明できること，さらには，雑音環境下でコミュニケーションの学習をすると連続値信号が 2 値化することも示した。

Google DeepMind は，ニューラルネットに convolution の構造を導入し，ATARI のゲームの強化学習を行なった [14]。そして，単にゲーム画面を入力し，ゲームスコアから求めた報酬からジョイスティックの動かし方を学習するだけで，ゲームの戦略という非常に知的な能力が獲得された。この大きなギャップは本アプローチの有用性を顕著に示した例と言える。

さらにわれわれは，リカレントニューラルネットを用いることで，注意，予測，交渉，運動制御，探索などの記憶を必要とする様々な機能が創発することを示した。このときは，強化学習に基づいて自動生成した教師信号を元に，BPTT (BackPropagation Through Time) 法を用いて時間を遡って学習を行った。これによって,「何を記憶すべきか」も学習を通して把握し，それを自分の行動に活かすまで一貫して学習することができた。

一般に BPTT によるリカレントネットの学習は誤差消失問題などで難しいとされ，現在は LSTM[15] などの方法が良く使われる。しかしわれわれは，一般的なリカレントネットを用い，状態遷移行列が単位行列に近い形になるように初期重み値を設定することで，誤差信号の過去への効率的な伝搬と双安定状態の容易な形成を実現し，記憶が必要な問題を学習させてきた。

5　探索も頭の中で作り出そう

「高次機能」の実現には，記憶したり考えたりと，ニューラルネットの内部でダイナミクスを操ることが重要となる。そのため，リカレントネットを使って強化学習をさせてきた。当初は「記憶」が必要なタスクの学習に重点を置いたため，前述のように，収束を重視した初期重み値の設定をしてきた。しかしこれでは，外部から入力が入らないと内部ダイナミクスは変化しなくなる。また，実際に学習させると，外部入力によって内部状態を遷移させることすら簡単ではなかった [16]。

一方で，ニューラルネットを用いて強化学習を行う際，Q-learning を使う場合は，ニューラルネットが求めた Q 値を元に確率的な行動選択を行い，連続値動作が可能な Actor-Critic を使う場合は，ニューラルネットが出力した動作信号に対して，それを中心とした確率分布から動作選択を行う。つまり，いずれも図 6(a) のように，サイコロを振る過程は一番最後に限られており，そこまでの動作生成とは切り離して考えられてきた。

End-to-End の考え方で，モータや筋肉への動作信号を直接ニューラルネットが生成するとすれば，その探索は，いわば手足で探索するようなものである。しかし，われわれにとっての探索は，たとえば分かれ道でどっちに行こうかと迷う場合は，手足のレベルではなく，脳の内部で行われているように感じられる。つまり悩んだ上での動作生成であり，最後の最後にサイコロを振るようなものとは明らかに異なる。そこで，図 6 のように，探索を強化学習のための特別な手段と捉えるのではなく，他の機能と同様にニューラルネット内で動作を生成するための処理の一部と捉えるべきと考えてきた。

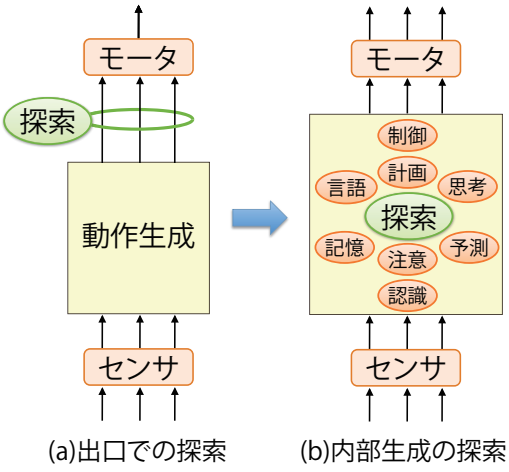

図 6: 探索の内部生成による動作生成と探索の一体化

6　「思考の創発」という夢に向けた仮説と実現の鍵となるカオスニューラルネット

われわれが難しい問題を「考える」ときには，あれこれと「迷う」。この「考える」ことと「迷う」ことは切っても切れないものである。また，われわれは学んだことを反映して「考える」ことができるようになる。そこで著者は，図 7 のように，『内部ダイナミクスとして生成される探索は，学習することによって状態遷移が合理的，論理的になり，その結果として「思考」が創発する』，しかし，『未知の状況に置かれると再び探索を行うようになる』という新しい仮説を提唱してきた。

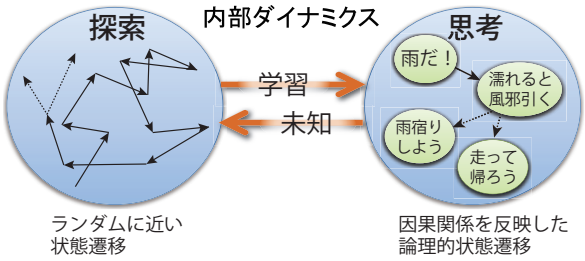

図 7: 学習による探索から思考への成長の仮説

それを実現する鍵となるのがカオスニューラルネットである。カオスネットは，カオスダイナミクスを生成するリカレントネットのことであり，ある程度大きいランダムな相互結合を設けたり，不応性を持つカオスニューロンを導入したり [17] することで実現できる。カオスダイナミクスは，小さな差が後に大きな違いを生み出すため，ランダムに近い動きを生み出すことができ，探索に利用することができる。さらに，カオスネットを用いた連想記憶における「カオス的遍歴」が，埋め込んだ記憶を自ら順次想起していく過程を観察すると，ものを「考える」ために必要なダイナミクスであると感じられる。

もう少し整理すると，図 8 のようなイメージになる。カオスダイナミクスを生成しない普通のリカレントネットでは，重み値が小さいとあまり大きな継続的な内部状態の変動を起こさない (R-0)。しかし，記憶が必要な学習を行うと，その記憶パターンを固定点とするアトラクタが形成され，固定点収束のダイナミクスを持つ連想記憶が形成される (R-1)。連想記憶は，記憶パターンを直接ネットワークに埋め込む形が一般的であるが，前述のリカレントネットを用いた強化学習の枠組みで，タスクに必要な情報の抽出を記憶と同時に学習させ，連想記憶が創発することも確認している [18][19]。一方，前述のように，自律的に状態を遷移させるようなフロー型のアトラクタを学習によって形成することは困難である (R-2)。

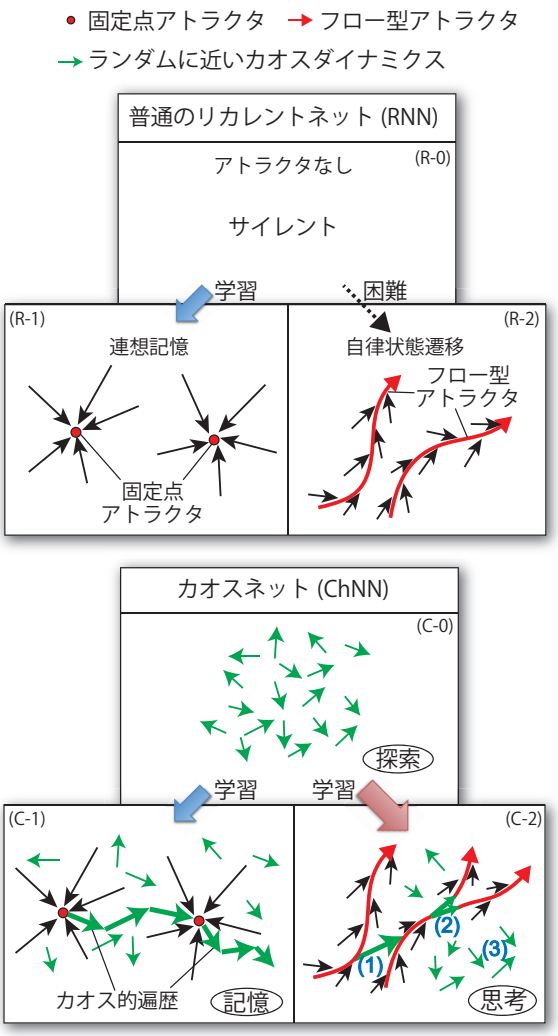

図 8: 普通のリカレントネットとカオスネットの内部での学習による 2 種類のアトラクタの形成と思考の創発

これに対し，カオスネットでは，ランダムに近い探索と呼べるダイナミクスを生成する (C-0)。ここに記憶パターンを埋め込むことでカオスダイナミクスの中に固定点アトラクタが形成され，前述のカオス的遍歴が観察される (C-1)。本論文の最初に述べた多義図形の認知過程も，これに近い形ものと考えられる。

本題の「思考」については，このカオスネットを学習させ，自律的な状態遷移を行うフロー型のアトラクタが形成されることで創発すると考えている (C-2)。このフロー型のアトラクタによる論理的，合理的な内部状態遷移を学習によって形成することはまだできていない。しかし，カオスネットは非常にリッチなダイナミクスを生み出すため，出力部分だけの学習で複雑な時系列パターンを容易に生成できる [20][21]。このことから，一般的なリカレントネットのサイレントな状態からスタートするよりは学習がはるかに容易になると期待している。

また図 8 の右下の図のように，カオスダイナミクスの上にアトラクタを形成することにより,「思考」に必要と考えられる 3 つのダイナミクス

(1) ひらめき or 発見
　　カオス的遍歴に近い形で，他のアトラクタが近くに来るとそのアトラクタに不意に移動する

(2) 高次探索
　　分かれ道での選択のように，可能な選択肢が複数ある場合に，ランダムに近い形で選択をする

(3) 探索への回帰
　　未知の状態になると，近傍にアトラクタが存在しないため，探索状態に回帰する

が創発することを期待している。

7　壮大な夢に向けて

思考は究極の高次機能であり，壮大な目標である。「何ができれば思考ができたと言えるのか」の基準は人それぞれで大きく異なるであろう。しかし，合理的な自律的状態遷移および前述の 3 つの必要条件についてはある程度コンセンサスを得られるのではないかと考えている。この考え方が今後の高次機能の研究を大きく転換し，新たな可能性を切り開く基盤となることを期待している。

カオスニューラルネットを用いた強化学習では動作信号が探索成分と一体となって生成されるため，探索成分と動作信号の分離ができない。そのため，探索成分の分離を前提とした従来の強化学習をそのまま適用することはできず，新たな学習方法の開発が必要であった。

筆者らは，各ニューロンの結合ごとに，出力の時間変化を見ながら過去の入力を効率的に保持する因果トレースと TD 誤差から学習する方法を提案した [22]。そして，中間層にリカレント構造を導入してカオスダイナミクスを生成するカオスネットを用いて学習をさせて来た。そして，カオスネットを用い，外部から探索成分を与えない全く新しい強化学習で，簡単なロボットの物体到達タスクを学習できることを示した。しかしながら，現時点では内部ダイナミクスの形成に重要と考えているリカレント部分の重み値の学習は機能しておらず,「思考」と呼べるダイナミクスは創発していない。

前述の 3 つのダイナミクスについては，学習途中でタスクの条件を変更することで，(3) の学習者の行動が探索的になることを示した [22][23]。また，この際，学習している環境については，学習とともにリアプノフ指数が減少する一方で，学習していない環境に対するリアプノフ指数はあまり減少しないことも確認した [22]。

(2) の高次探索としては，1 次元視覚センサを用いた車輪ロボットによる障害物回避問題を学習させた結果，障害物の右にいるときは右側を，左にいると左側を通ってゴールに向かうが，障害物正面では，障害物の右を通るか左を通るかをランダムに近い形で選択されることを示した [24]。ただし，まだ障害物に衝突することもあり，改善の余地を残している。

今後は，強力な学習則であり，機能創発に力を発揮してきた誤差逆伝播 (BP) 法との併用を探り，リカレント部の重み値の学習を目指したいと考えている。そして，壮大な目標である「思考」の創発に今後も引き続きチャレンジしていきたい。

【本論文の内容に関する詳細は文献 [10] を，個別の機能の創発については文献 [10] 中の参考文献をご覧ください。】

謝辞

本研究は JPSP 科研費 15K003600 の補助を受けた。研究室の皆さんとこれまで受けた補助に謝意を表します。

参考文献

[1] K. Shibata & Y. Okabe: Reinforcement Learning When Visual Signals are Directly Given as Inputs; Proc. of ICNN(Int'l Conf. on Neural Networks)97, Vo. 3, pp. 1716-1720 (1997)

[2] 柴田克成: ニュー ラルネットワークを用いた Direct-Vision-Based 強化学習 ーセンサからモータまでー; 計測自動制御学会論文集, Vol. 37, No. 2, pp. 168-177 (2001)

[3] 理研脳科学総合研究センター: ニューロン; http://www.brain.riken.jp/jp/aware/neurons.html (2018.3.5 現在)

[4] R. Levi-Montalcini: Developmental neurobiology and the natural history of nerve growth factor; Annu. Rev. Neurosci. Vol. 5, pp. 341- 362 (1982)

[5] 津本忠治: 脳と発達; 朝倉書店 (1986)

[6] Y. Murata, et al.: Temporal Plasticity Involved in Recovery from Manual Dexterity Deficit after Motor Cortex Lesion in Macaque Monkeys; J. of Neuroscience, Vol. 35, No. 1, pp. 84-95 (2015)

[7] A. Krizhevsky, et al.: ImageNet Classification with Deep Convolutional Neural Networks; in Advances in NIPS, Vol. 25, pp. 1097-1105 (2012)

[8] D. Dennett: Cognitive Wheels: The Frame Problem of AI. Minds; Machines and Evolution, Cambridge Univ. Press, pp. 129-151 (1984)

[9] B. Goertzel.: 汎用人工知能概論; 人工知能学会誌, Vol. 29, No. 3, pp. 228-233 (2014)

[10] 柴田克成，後藤祐樹: 深層学習が示唆する end-to-end 強化学習に基づく機能創発アプローチの重要性と思考の創発に向けたカオスニューラルネットを用いた新しい強化学習; 認知科学, Vol. 24, No. 1, pp. 96-117 (2017)

[11] D. Hassabis: Google DeepMind - Artificial Intelligence & the Future; Seminar at KAIST, Korea, https://youtu.be/8Z2eLTSCuBk, Mar. 11 (2016)

[12] K. Shibata & T. Kawano: Learning of Action Generation from Raw Camera Images in a Real-World-like Environment by Simple Coupling of Reinforcement Learning and a Neural Network. Adv. in Neuro-Information Processing, LNCS, 5506, 755-762 (2009)

[13] R.S. Sutton & A.G. Barto: Reinforce Learning, A Bradford Book, The MIT Press (1998)

[14] V. Mnih, et al.: Human-level control through deep reinforcement learning; Nature, Vol. 518, pp. 529-533 (2015)

[15] S. Hochreiter & J. Schmidhuber: Long short-term memory; Neural Computation, Vol. 9, No. 8, pp. 1735-1780 (1997)

[16] Y. Sawatsubashi, et al.: Emergence of Discrete and Abstract State Representation through Reinforcement Learning in a Continuous Input Task; Adv. in Intelli. Sys. and Comp., Robot Intelli. Tech. and Appli. (RiTA) 2012, pp. 13-22 (2012)

[17] G. Matsumoto, et al.: Periodic and Nonperiodic Responses of Membrane Potentials in Squid Giant Axons During Sinusoidal Current Stimulation; J. of Theoretical Neurobiology, Vol. 3, No. 1, pp. 1-14 (1984)

[18] K. Shibata & M. Sugisaka: Dynamics of a Recurrent Neural Network Acquired through Learning of a Context-based Attention Task; Artificial Life and Robotics, Vol. 7, No. 4, pp. 145-150 (2004)

[19] K. Shibata & H. Utsunomiya: Discovery of Pattern Meaning from Delayed Rewards by Reinforcement Learning with a Recurrent Neural Network; Proc. of IJCNN 2011, pp. 1445-1452 (2011)

[20] H. Jaeger. : The "echo state" approach to analysing and training recurrent neural networks. German National Research Center for Information Technology GMD Technical Report 148.34 (2001)

[21] W. Maass, et al.: Realtime computing without stable states; Neural Computation, Vol 14, No. 11, pp. 2531-2560 (2002)

[22] 柴田克成，坂下悠太: カオスニューラルネットを用いた内部ダイナミクス由来の探索に基づく強化学習; 電子情報通信学会技術報告, NC2014-117, pp. 277-282 (2015)

[23] T. Matsuki & K. Shibata: Reward-Based Learning of a Memory-Required Task Based on the Internal Dynamics of a Chaotic Neural Network; Proc. of ICONIP 2016, LNCS 9947, pp. 376-383 (2016)

[24] Y. Goto & K. Shibata: Influence of the Chaotic Property on Reinforcement Learning Using a Chaotic Neural Network; Proc. of ICONIP 2017, LNCS 10634, pp. 759-767 (2017)