特　集「人工知能学会・情報処理学会共同企画―第３部「技術紹介」―」

汎用性の創発を脳に学ぶために

How to Introduce General Intelligence of Brain?

山川　宏　株式会社ドワンゴドワンゴ人工知能研究所
Horoshi Yamakawa　Dwango Artificial Intelligence Laboratory.
http://ailab.dwango.co.jp/

1. 表 現 と 推 論

深層学習の技術的進展が近年の第三次人工知能（AI）ブームを牽引してきた．この流れを引き継いでAIを今後発展させていく方向性として，汎用AIや脳型AIがある．いずれも以前から存在する研究アプローチではあるが，深層学習技術を取り込むことで大きく前進し，その間の関係性も様変わりした．

本稿では，AI技術の基本的な構成要素である，表現と推論という二つの側面を軸に著者なりの視点から本動向を俯瞰する形での解説を試みたい（図1）．

1·1 記号表現と分散表現

AIにおいて知識の記述に用いられる表現は大きく，記号表現と分散表現に分けられる．

- ●記号表現：局所的な値に特定の意味（セマンティクス）を対応付けた表現．自然言語における単語などに相当する．コミュニケーションやプログラミングに利用できる．
- ●分散表現：多次元ベクトルなどにおける値の分布を用いて概念を表現する．通常は外部から意味付けしないため，コミュニケーションやプログラミングへの利用は困難だが，多次元空間中で概念間の距離や関係などを用いた処理は可能である．

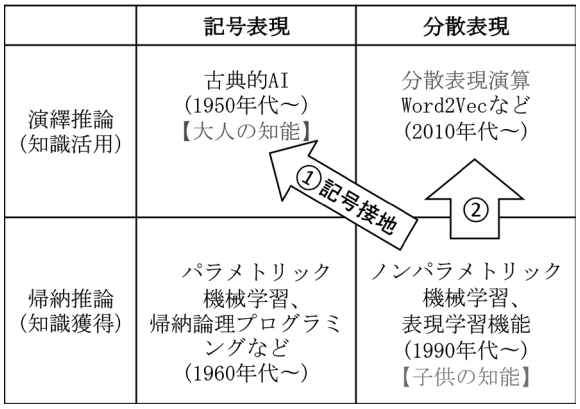

図1　AI技術における表現と推論．
図中で①と②は分散表現による知識に対して演繹推論を行う二つの方法を示している

1·2 帰納推論と演繹推論

AIに用いられる論理的推論は，主には帰納推論と演繹推論であり（本稿ではアブダクションは議論しない）[米盛07]，AIの観点からは以下のように捉えられる．

- ●帰納推論：データから知識を獲得する推論．典型的な機械学習．
- ●演繹推論：知識を組み合わせて活用する推論．データがない場合や少ない場合にも利用できる．

表現と推論の両軸からAI技術を俯瞰しつつ（図1），おのおのについて説明する．

1·3 知識を活用する演繹推論：分散表現演算（DROP）技術の勃興

AI研究における古典的な流れは，記号表現を用いた演繹推論であり，1950年代から始まっている．これは小学生以上で発達する計画や論理思考といった，いわゆる大人の知能である．

これに対し，分散表現上の知識を直接活用しての演繹推論を行う技術を新たに分散表現演算（Distributed Representation Operation：DROP）と呼ぶことにする．つまり分散表現演算それ自体は，データからの学習ではなく獲得した知識の活用である．よって，分散表現演算を機械学習の概念に含めるか否かは少々難しい判断となるが，実体として機械学習の研究コミュニティが主導している．

深層学習の進展を背景に近年，この種の研究が盛んになってきた．例えばWord2vecにおいては言語情報を分散表現上に変換したうえで類推する機能がある．また，Generative Adversarial Netsは，静止画像データから物体の知識を獲得し直接には学習していない画像を生成する機能が実現されている．

1·4 帰納推論による知識獲得と事前知識

機械学習器は，データから知識を獲得する帰納推論のための装置である．一般的に機械学習器はデータを利用して数学的なモデルに含まれるパラメータを調整することで知識を獲得する．よって機械学習器が有する知識は，設計者が事前にモデルに組み込んだ知識と，データから獲得した知識であり，定性的には，それら知識の総和に

より性能の上限が規定される．

機械学習器が対象とするタスク範囲が狭ければ，より多くの事前知識をモデルに組み込み得るし，そこに含まれるデータの性質もそろうので，良い性能を得やすい．逆にタスクの範囲が広くなると事前に共通的に設計できる知識が少なくなり，性能を上げることが難しくなる．

機械学習が発展してきた歴史を振り返ると，基本的には，計算量とデータの増大に伴って，設計する知識よりも学習で得る知識へのウェイトが増している．

初期の機械学習の研究は，線形回帰モデルのように明示的に意味付けし得るパラメータの調整から始まっている．当然ながらこうした手法では，設計時の想定を超えた表現は獲得されない．

設計時にパラメータの意味は決定しなければ，潜在的には学習で獲得した表現の一部分については外部からの観測で意味付けし得る可能性がある．その後に発展した人工ニューラルネットワーク（ANN）などのノンパラメトリック機械学習がこれにあたるが，深層学習の発展を経てそれは現実のものとなった．これは乳児が認知発達を通じて暗黙的に獲得する知能の性質に合致することから，ここではこれを子供の知能と呼ぶことにする．

近年の深層学習の成功は [LeCun 15]，基本的にはANN 技術に対して高速計算と大規模データが適用できる環境が整った結果である．重要なことは「十分にデータを得られるタスクの範囲内であれば，応用価値のある人間並みの性能をもつ帰納推論が可能になった」という点である．つまり計算パワーの増大により，最近ようやく子供の知能が実現されたのである．

こうして，図 2 に示すように大人の AI と子供の AI が出そろうことで，その統合を通じた人間のような AI の実現に向けた動きが本格化し始めているのである．

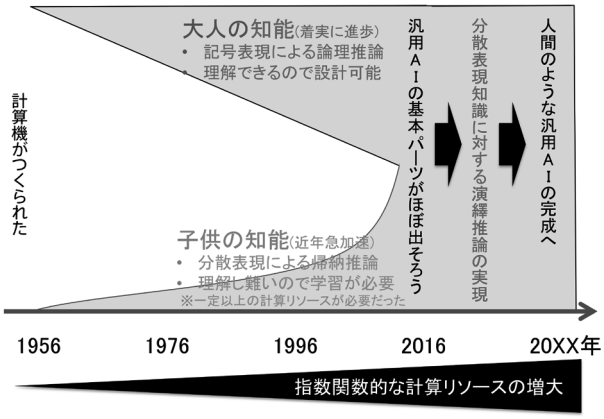

図 2　大人の知能と子供の知能の発展と合流．
2015 年頃から二つの知能が合流しはじめたことで，
汎用 AI 研究が本格化しはじめた

2. 人型 AI と脳型 AI

AI の研究開発を進める際の目標設定として，人の振舞いもしくは人の脳に似せてつくるか否かは重要な判断軸である．

そこでまず AI 一般，人型 AI，脳型 AI を，研究目標の面から整理すると以下のようになる．

- ●AI 一般の目標：産業や科学技術開発などの応用を志向する，何らかの意味でより高度な知的能力の実現
- ●人型 AI に固有の目標：人のように振る舞う知能の実現
- ●脳型 AI に固有の目標：脳の機能的な理解に基づき医療貢献，さらにはマインドアップロードなども目指す

AI に期待される社会貢献や応用シーンでは人や脳に固有な知能は必ずしも必要とされない．一方で，AI 研究は常に人の知能との比較をしてきた．こうした関心が高いのは，あるタスクにおいて AI の知能が人間を上回れば，それに関わる職業は AI に代替される可能性が高いからであろう．

さらに脳型 AI では脳メカニズムから医療に貢献する側面なども研究目標に含まれ得る．

AI 一般を目標とした場合には，人型 AI や脳型 AI を目指すことは到達手段になる．同様に人型 AI を目標とした場合には脳型 AI は到達手段になる．脳型 AI において脳全体を扱うのであれば，人型 AI は一つの目標となり得る．

3. 汎　　用　　AI

3·1　汎用性という技術目標

汎用 AI は，現在実用化されている特化型 AI に対置して現れた概念であり，2006 年頃に Ben Goerzel 氏により提唱された．大雑把に言えば多種多様な課題に対して問題解決を行える AI を構築しようとする試みである．逆に現在実用的な特化型 AI は，対象とするタスク（例えば，囲碁，自動運転など）に応じた事前知識が潤沢に組み込まれている．

AI 研究の目標として汎用性が着目される理由に以下があるだろう．

- ●人間が備えている知能
- ●現状の AI で実現されていない知能
- ●性能評価できる知能（実は難しいのだが）

汎用 AI の性能評価に対する指標は主に 2 種類ある．先の AI 一般を目標とする立場では「経験からの学習を通じて，さまざまな問題に対する多角的な解決能力を獲得できる知能」となる．対して人型 AI の立場では，人と同程度に多種多様な知的能力を発揮できるという側面から評価される．

逆に「汎用 AI は何ではないか」を指摘するなら，以下のようにいえるだろう．

- ●単に特化型 AI の寄せ集めではない

●最初から何でもできる知能ではない
●タブラ・ラサ（白紙）から学習するのではない
●意識の有無は考慮しない（評価が困難）

3·2　深層学習の後押しで本格化する汎用 AI 研究

個別のタスクに着目すれば，大量データが得られる状況であればANN などの帰納推論が人を超えた知的能力を実現できることも多い．

こうした背景から，最近の汎用 AI の国際会議などにおいて議論される AI の基本課題は個別のタスクを離れ，主に以下のような側面が着目されている．

●汎用性
●現実的な時間内での問題解決
●少数データへの対応
●演繹推論（計画などの大人の知能）

実はこの四課題は図 3 に示すように深く関係する．知能の汎用化を目指してタスクの適用範囲を広げていけば，しばしばデータ不足が生ずる．ここで現時的な時間内に問題解決を行おうとすれば，十分なデータを収集するほどの時間的余裕がない．こうしてデータ不足が解消できない状況では帰納推論のみでは良い性能は得られない．そこで，演繹推論を導入する必要が生ずる．こうして，今後の AI 研究開発における基本的課題として，汎用性を中心とした側面はより鮮鋭化したといえそうである．

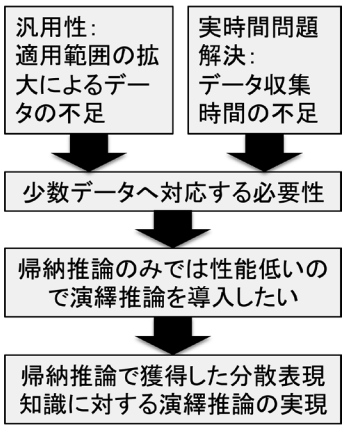

図 3　鮮鋭化する汎用性を巡る課題

しかしながら図 1 に示したように，ANN などの帰納推論では分散表現として知識が獲得される．しかし，伝統的な演繹推論は記号表現を用いて行われるため，単純に組み合わせることはできない．そこで「分散表現知識に対する演繹推論」の実現という新たな課題が生ずる．

これに対する解決法は以下の何れか，もしくはその両方の組合せとなるであろう（図 1）．

① 記号接地アプローチ：分散表現知識を記号的知識に対応付けたうえで記号により演繹推論などを行う．
② 分散表現演算アプローチ：分散表現知識に対して直接に演繹推論などを行う．

4. 脳型 AI：脳を参照して汎用性に迫る

脳型 AI の基本的な目的は AI をつくることにある．より確立した近隣分野として計算論的神経科学 [銅谷 07] の分野が存在するが，こちらはむしろ理解をするために計算論であり目的が異なるが，互いに参考になる学術領域である．

4·1　脳型 AI の実現性が高まった

現存する汎用 AI は脳以外にはなく，第一義的にはこれをまねて汎用 AI を構築する脳型 AI のアプローチは早道かつ自然に思える．ANN は脳の神経回路を模したものであるし，深層学習は脳の視覚野の階層構造を参考にしている．また大脳基底核の機能は強化学習との対応がよくとられている．全般的に見ても図 4 に示すように脳に対応付ける形で，AI の基本的な仕組みを対応させることは可能である [WB AI]．

これまでは脳型 AI を推進するにあたり二つの大きな課題があったが，解決されつつある．

一つ目に，汎用 AI の特性を生み出すうえで重要な役割を担う大脳新皮質に対して，工学的に有用な情報処理を行える計算モデルが存在しなかった．それをある程度模倣できる形で深層学習が現れた点があげられる．ここで新皮質は，その全領野にわたりほぼ共通の 6 層構造をもつ局所回路で構成されるが，入出力される情報に応じ

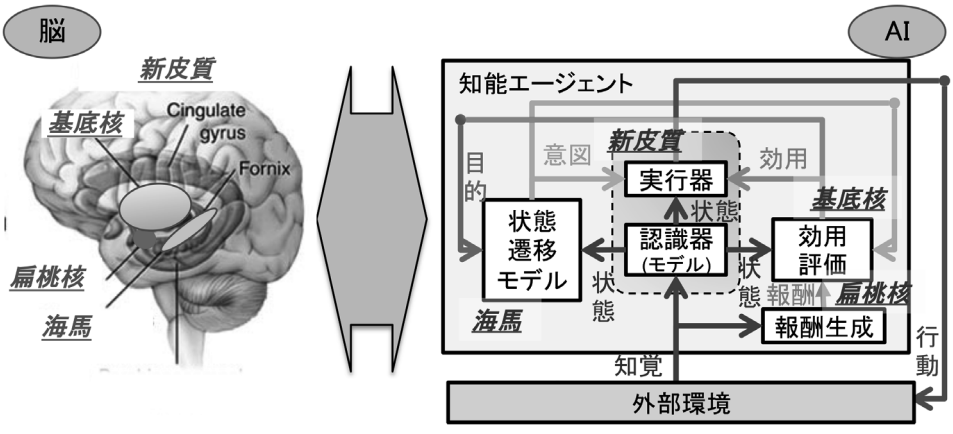

図 4　脳に対応付けられる AI

て異なる機能を実現する [山川 14]．これが学習によって多様な機能を獲得する汎用 AI の特性に対応する．

また神経科学知見は，必ずしも AI の発展に大きく役立ってきたわけではない．その大きな原因は，従来の神経科学知見は脳全体のマクロな振舞いと，神経細胞数個のミクロな振舞いへの理解が切り離された形で進展するというミクロ–マクロギャップが存在したためである．しかし現在は，光遺伝学の進歩により動物実験では，ある部位における 1 000 個規模のニューロン活動を同時計測が可能になり，その神経活動を制御できるようになった．また脳全体の静的なネットワーク構造や，局所的な詳細なネットワーク構造がコネクトーム研究で明らかにされつつある [スン 15]．さらに脳内で空間的に離れた複数領域の活動を同時測定し，その関係を明らかにするさまざまな研究も進んでいる．つまり近年，神経科学の研究は階層的なネットワークの動的な理解に立ち入りつつあり，ようやく AI の設計にとって有用な段階となってきた．

4·2 汎用 AI 開発において脳を参照する意義

ストレートに脳の仕組みを同定して，それを計算機上に実装することで脳型AIがつくれるならわかりやすい．しかし，そこまでの脳の理解が進むにはいまだ長い年月を要しそうである． そこで現状の，脳型 AI のとるスタンスは，脳はあくまで参照として用いるしかない．

そのうえで脳型 AI のアプローチをとるメリットには二つの方向がある．

【A】（人間には難しい）設計のガイド

- ●現状の AI で実現されていない未解決な計算機能についてのヒント（暫定モデル）が得られる．
- ●目的が多岐にわたる汎用 AI は，目的に応じた機能分割による設計が困難であることに対処．

【B】分散共同開発の足場

- ●開発初期段階から脳の機能と構造に合致するように実装を制約し，後々に技術統合の問題が生じるリスクを回避する．
- ●周辺の科学的知見（認知科学・神経科学など）を集約する足場となる．
- ●汎用 AI に必要な機能の抜けを確認できる．
- ●チーム内で「脳に近い実装を優先する」という価値観を共有すれば，大規模分散開発の発散を防止できる．

4·3 いかなる粒度で脳を参照すべきか

脳は階層的なネットワーク粒度をもつため，どのレベルの粒度までを参照しながら，汎用 AI の実現を目指すかは，それを最も早期に完成するという点からは重要な選択となる．より細粒度でモデル化すれば，以下のメリットがある．

- ●人の脳により近い形で動作すれば，汎用 AI に到達できる可能性が高まる
- ●神経科学知見を直接的に利用しやすい

逆にデメリットとしては以下がある．

- ●パラメータが膨大になりその決定が困難になる
- ●実行や学習の計算量が増大する

抽象度の高いモデルではこのメリット・デメリットは反転する．

十分な知的機能を備えた脳型 AI を実現するために，脳をどこまで詳細に再現すればよいかは，実際にできてみるまでわからない可能性が高い．よって，この経路から汎用 AI を最速で実現するためには，各時代において利用し得る計算リソースのもとで，学習を含めて試み得る詳細さと規模をもつモデルを試すことになる．時代を経てさらに計算リソースが増大すればモデルをより詳細化・大規模化していけるだろう．

同じ物体を見せた，脳内で物体認識を行う視覚野の神経活動と，物体認識を行うように学習した深層学習モデルの振舞いが良く対応付くことが示されている [Cichy 16]．つまり少なくとも視覚野においては「学習結果として得られる ANN の動作については，脳にかなり近い」といえる．

現在の ANN に多用されるニューロンモデルはかなり単純なものである．具体的には，連続値による多入力 1 出力の非線形関数で，内部構造をもたない（マカロック・ピッツモデルの延長線上）．また，神経科学では標準的なスパイクとしての挙動も平均化されている（rate coding）．

つまり脳型 AI をつくる目的においては，ここまで単純化されたニューロンモデルを仮定しても十分である可能性が高まったのである．

ただし学習アルゴリズムについては，既存の ANN で多用される誤差逆伝搬法の生物学的な妥当性が低く，その点についてはより詳細に生物に似せたモデル化が必要となる可能性はある．

4·4 価値を司る回路

AI における意思決定において，とり得る行動選択肢の価値を評価し，その値が高いものをより多く選ぶ方法は代表的なものである．これは効用エージェントとも呼ばれる．

選択肢の効用を評価する機構を価値システムと呼べば，さらに以下二つのサブシステムに分解される（図 4）．一つ目は，外部環境（エージェントの身体も含む）から得られる状態や行動についての報酬（価値）を決定する報酬生成部である．これは AI 自身にとっての目標を決定するものである．二つ目は，上記の報酬から状態や行動についての間接的な価値である効用を計算する効用評価モデルである．これは AI においては強化学習アルゴリズムで実装される．

人を含む動物において，状態や行動に対する価値評価

は感情と呼ばれる．快・不快と直結した一時的で急激な感情は情動と呼ばれ，例えば，怒り，恐れ，喜び，悲しみなどがある．これ以外の感情として例えば，諦め，驚き，嫌悪，恐怖，不安，共感，嫉妬などがある．

価値システムの神経科学的基盤はいまだ明らかにされたとは言いがたいが，AIのサブシステムとはおおむね以下のように対応する．

●報酬生成部：間脳，脳幹，扁桃核，など

●効用評価モデル：主に大脳基底核

報酬生成部は，情動との関連が深く，動物が自身の生存の起点となる情報の価値付け（食欲など）を行う．関連する脳器官は進化的に古いため，その神経回路は概ね遺伝的に決定されており，学習で変化する要素は少ない．それゆえ，複雑な神経回路網を個別的に理解する必要があり未解明部分が多い．対照的に，新皮質，小脳，大脳基底核，海馬といった比較的新しい脳器官は比較的均一な構造が学習を通じて変化することで機能を獲得する．

もう一方の，効用評価モデルは，大脳新皮質と再帰的に連結した大脳基底核により脳内に実装され，主に強化学習アルゴリズムをベースとした意思決定に関与する．さらにエピソード記憶などを扱う海馬とも連携している．こうした大域的な回路により，共感，妬み，尊敬などの複雑で豊かな感情を生み出している [Koelscha 15]．

AIにおける価値システムの存在意義を考える．特化型AIは，報酬生成部をもたせずにユーザがエージェントの外部から直接に報酬信号を与えるか，設計者が目指すタスクに応じた報酬を，報酬生成部内に設計する．こうした，特化型AIにおいても，中間的な価値を設計する効用評価モデルの役割は大きい．例として，ビデオゲームでの好成績を目指すAIでは，深層学習に効用評価モデルとしての強化学習を組み合わせている．また，膨大な文献を理解しつつ物理学実験の計画を立てるAIにおいても研究計画の中間目標を設定することは有益であろう．こうした，道具として用いられる特化型AIの範囲においては，価値システムは基本的にユーザや設計者による管理の影響が大きく，その振舞いが人間の情動や感情と同質であることはあまり想定されない．

一方で，人と共存する人型AIにおいては，人の気持ちを理解したり人のように振る舞ったりすることを目指す．そこで人間と同様に，さまざまな刺激に対して価値付けを行う報酬生成部を設計することになる．

なお今後出現する汎用AIの典型的なイメージは，家事ロボットであろう．こうした汎用AIにおいては，複数の基本目標の達成を目指しつつ，安全性を確保するための制約なども考慮することになる．こうした状況において，効率的に意思決定を行うために，さまざまなコンフリクトを解消しながら中間目標を設定するような高い自律性を伴う価値システムが必要になるであろう．

4·5 分散表現に対する演繹推論は脳でいかに創発しているか

すでに3章で述べたように，今後の汎用AIにおいて，「分散表現知識に対する演繹推論」は大きな課題である（図1）．ここでは上記の機能が脳においてどのように実現し得るか検討する．

一つ目（①）は，長年にわたるAIの課題とされてきた記号接地問題に関わり，主な関連分野として，記号創発ロボティクスなどがある．こうした分野では機械学習を活用することで複数のモダリティーの信号を言語概念に結び付ける技術などに進展がある．神経回路からのモデル化としては，大脳新皮質上において話す機能を担う領野（ブローカ野）や聞く機能を担う領野（ウェルニッケ野）を中心としたモデル化研究が進展している [Kemmerer 15]．

脳内において，二つ目（②）の分散表現演算に用いられる知識を考える際には，「ユニークネス」という性質を仮定すべきであろう．つまりある知識は神経ネットワーク上のある部分領域に存在し，それ以外の場所に同じ知識は存在しないという性質である（例えば，ネコという概念を表すのと同一の知識の本体は複数のネットワークには存在しない）．なぜなら知識は学習を通じて変化し続けるうえに，脳内においては知識を複写する操作はないからである．

こうしたユニークネスの制約から，複数の分散表現知識を組み合わせて，分散表現演算を行うためには，基本的には直接的に知識を相互作用させる必要があると思われる．

具体的に神経科学の面から研究が進んでいる例として，げっ歯類（ネズミなど）の海馬および嗅内皮質における空間推論がある．彼らは，自己の移動した経路を考慮してショートカットによる帰巣経路を計算するような一種の演繹推論が可能である．この場合には，海馬周辺に集約された空間に関わる知識を連携させる形で分散表現演算が実行されているように見える．

また，人間で特に発達した大脳新皮質上において，異なる領野ごとに多様で独自の知識が蓄積されていると想定される．そしてこうした異種の知識が大域的な神経ネットワークを通じて連携することで高度な分散表現演算が創発していると思われる．そこで今後さらなる進展が期待される神経科学分野から，学習で獲得した知識を柔軟に再利用する汎用AIを実現するためのヒントが得られると期待している．

5. お わ り に

本稿では，まず，AIの研究アプローチはどういったレベルで人間のような知能を目指すかという観点から整理した．人型AIは人のような振舞いを目指し，脳型AIは脳のメカニズムを参照する．これに対して汎用AIは，

現状のAIが個別のタスクに特化しているという限界を乗り超えようとする技術目標であり，その中でも，どの程度に人間のような知能を目指すかについては異なる立場が存在する．

次に，深層学習の進展がAI研究に与えたインパクトを概観した．それは「十分にデータを得られるタスクの範囲内であれば，応用価値のある人間並みの性能をもつ帰納推論（機械学習）が可能になった」といえそうである．ところが，汎用AIは経験から得られた知識を組み合わせて，データが少ない新たな状況にも対応する必要がある．原理的には演繹推論を組み合わせることで，データから得られた分散表現上での知識を活用したい．しかし伝統的な演繹技術は記号表現を対象としているため，そのままでは分散表現に適用できない．こうして「分散表現知識に対する演繹推論」が重要な研究トピックになりつつある．こうした側面は，以前より繰り返しAI分野で議論されてきたが，深層学習の進展を背景に，汎用AIという研究コンテクストにおいて鮮明化している．

この課題を解決するためのアプローチは，主に二つであろう．一つ目の方法は，分散表現の知識を記号に接続したうえで記号的に演繹を行うアプローチである．二つ目は，分散表現上の知識を直接に活用する分散表現演算とでも呼ぶべきアプローチであり，機械学習分野から派生しはじめている．

神経科学の急激な進展により，知能について脳からヒントを得て，脳型AIの構築を進められる状況になってきた．さらに人間の脳における知識は基本的には分散表現として蓄えられており，どちらかというと記号による演繹推論は不得意である．そう考えると脳における演繹推論はむしろ分散表現演算が主体である可能性すらある．こうした考察から，著者としては特に，今後において知能を汎用化することに役立つであろう分散表現演算技術は，脳からヒントを学び得ると考えている．

◇ 参 考 文 献 ◇

[Cichy 16] Cichy, R. M., et al.: Comparison of deep neural networks to spatio-temporal cortical dynamics of human visual object recognition reveals hierarchical correspondence, *Scientific Reports*, Vol. 6, Article number: 27755（2016）

[銅谷 07] 銅谷賢治：「計算神経科学への招待」脳の学習機構の理解を目指して，臨時別冊数理科学，2007年12月号（2007）

[Kemmerer 15] Kemmerer, D.: *Cognitive Neuroscience of Language,* Psychology Press（2015）

[Koelscha 15] Koelscha, S., et al.: The quartet theory of human emotions: An integrative and neurofunctional model, *Phys. of Life Reviews,* Vol. 13, pp. 1–27（June 2015）

[LeCun 15] LeCun, Y., Bengio, Y. and Hinton, G.: Deep learning, *Nature,* Vol. 521, pp. 436-444（May 2015）

[スン 15] セバスチャン・スン：コネクトーム：脳の配線はどのように「わたし」をつくり出すのか，草思社（2015）

[山川 14] 山川 宏，市瀬龍太郎，井上智洋：汎用人工知能が技術的特異点を巻き起こす，信学誌，Vol. 98, No. 3, pp. 238-243（2014）

[米盛 07] 米盛裕二：アブダクション―仮説と発見の論理，勁草書房（2007）

[WB AI] 全脳アーキテクチャとは：http://wba-initiative.org/wba/

2016年8月22日　受理

著 者 紹 介

山川　宏（正会員）

1989年東京大学大学院理工学系研究科物理修士課程修了．1992年同大学院工学系研究科電子博士課程修了．工学博士．同年，株式会社富士通研究所入社後，センサフュージョン，RWCプロジェクトに参加．現在，株式会社ドワンゴドワンゴ人工知能研究所所長．概念学習，認知アーキテクチャ，教育ゲーム，ニューロコンピューティングなどの研究に従事．電子情報通信学会，日本認知科学会，日本神経回路学会，日本テニス学会の各会員．本学会理事，2016年度から本誌編集委員長．