2603BrainSimilarAI.pdf
2026.3 ブログ：「脳と AI は似ているか」を読んで、の詳細
（→ http://www.1968start.com/M/blog/index4.html#2603 ）

「脳と AI は似ているか」を読んで

中所武司

■このエッセイのきっかけ
人工知能学会誌の最新号の下記の解説論文について、
私の卒論、修論を通じての学生時代から関心のある分野なのでコメントする。

・人工知能 Vol.41 No.2 (March 2026) pp.93-100
特集「AI と神経科学の接点 2026」
脳と AI は似ているか —NeuroAI の挑戦—

■内容の要約とコメント（→★）

1. はじめに

・「**脳と AI は似ているか**」という古くからの問いは，新たな局面を迎えている．
以前の AI は，心理学的知見や脳の仕組みをアルゴリズムへ実装するのが課題だった．
現代の**深層学習**のブレークスルーは，生物学的な詳細を厳密に模倣するよりも，
膨大なデータと計算資源を用いて性能を追求した結果としてもたらされた．

・脳を直接的に模倣するように設計されていない AI が，工学的な最適化の結果として
脳との類似性「アラインメント」を示すことがわかってきた．

・本稿では、AI と神経科学・心理学との歴史的な関係を概観し、
現代の深層ニューラルネットワーク（DNN）が脳の活動と対応し得る背景にある
「NeuroAI」という学際領域の理論的基盤を整理するとともに，
それが認知神経科学や心理学への依存をいかに乗り越えようとしているのかを論じ、
最後に，脳・AI・心を横断する共通の枠組みとしての「**潜在表現**」の概念に注目し，
主観的な心的事象を客観的な対象へと外在化する，新たな科学の展望を提示する．

2. 「苦い教訓」

・AI 研究の父の一人、H.A.Simon[文献 1983]は AI 研究に二つのゴールがあるとした。
第一は「**工学としての AI**」であり，人や社会に役立つ有能なシステムの開発を重視．
第二は「**理論心理学としての AI**」で人間の心的過程の理解とシミュレーションを目的．
前者は性能や利便性を，後者は心や脳の仕組みの理解とモデル化を重視する．

→★H.A.Simon は、人工知能に関する最初の会議（ダートマス会議,1956）に参加。

→★私は卒論では第一の目的、修論では第二の目的でニューラルネットワークを研究。
　卒論では、ニューロン間の結合を模した、のちの PLA のような回路を設計。
　修論では、学習によってニューロン間の結合係数が変化するモデルを作成し、
　コンピュータシミュレーションによって、討論学習などの実験を行った。

【参考：関連学会発表】
＊**条件反射における学習機能に注目した回路モデル**、電子通信学会全国大会、242(**1969**)
https://www.1968start.com/M/bio/olduniv/gakkai1969.html
＊**思考過程のシミュレ-ション**、電子通信学会オ-トマトン研究会、A70-76(Dec.**1970**)
https://www.1968start.com/M/bio/olduniv/gakkai7012.html

・AI 研究のもう一人の父、A.Newell[文献 **1973**]は，後者の立場から，
複雑なタスク全体を統合的に扱える「完全な処理モデル」の構築が重要と主張．
ただし，当時は計算資源や理論的知見が限られ，統合モデルの実現は困難だった．

→★A.Newell も、人工知能に関する最初の会議（ダートマス会議,**1956**）に参加。

→★私の修論では、計算モデルの Fortran プログラムを大型計算機センターで実行。
　CPU 時間 30 秒以内という制約のため、神経細胞間の結合係数の行列は 10*10 だった。

・現代の AI の源流の一つである強化学習のパイオニア，R.Sutton[文献 2019]は，
真のブレークスルーや飛躍的進歩は，計算資源のスケーリングと
大規模データによる探索・学習でもたらされてきたと強調．
実際，コンピュータチェス，囲碁，音声認識，コンピュータビジョンなどの分野で
人間を超えたのは，膨大な計算とデータによって自ら学習するアプローチだった．

・Sutton は，「実際の心の中身は途方もなく，複雑だ．心の中身について考える
簡単な方法を見つけようとするのはやめるべきだ」と主張．

→★耳の痛い話。私の修論では、出力を入力へフィードバックするモデルなので、
　多層ニューラルネットワークともいえるが、思考過程を 10*10 のニューロン結合で
　計算機シミュレーションして、学会発表した。(^^;;

3. 現代の AI＝DNN

・現代の AI の中核は，**深層ニューラルネットワーク**（DNN）である．
DNN はよく「脳やニューロンの構造・機能にインスパイアされている」と言われる．

→★DNN は、神経細胞間のシナプス結合に着目した構造ということかな。

・DNN に基づく AI の先駆者達（G. Hinton ほか）の一部が有する神経科学や心理学の
知見や成果が、AI のアルゴリズムや構造へ直接的に反映・応用されたとは言いがたい．
DNN はニューロンの構造と機能を極度に単純化したユニットのネットワークを
膨大なデータを用いた end-to-end 学習で訓練した計算モデルである．

→★G. Hinton は、AI の基礎を築いたという功績で 2024 年ノーベル物理学賞を受賞。

・古典的 AI の論理規則や知識表現を用いた「シンボリック」なアプローチとは対照的に，
DNN は人工ニューロンの活動パターンを通じて情報を暗黙的・分散的に表現するので、
内部表現は人間にとって直感的に解釈が困難な「ブラックボックス」と評される．

→★1980 年代の第二次 AI ブームでは、知識表現とそれを用いた推論方式が重要だった。

→★1960 年代の第一次 AI ブームでの私の修論の「古典的な AI」では、ニューロン A から
ニューロン B への結合度が大きい場合、概念 A から概念 B が連想されるとしていた。
DNN では、特定の AB の結合度が大きくても、その理由は説明できないので、
ある入力から特定の出力が得られた理由を、内部表現を使って説明できない。

ニューロン素子：

・**1943 年**に発表されたマカロック-ピッツ・モデルは，神経細胞の働きを抽象化して、
「入力の総和がしきい値を超えれば発火する」というシンプルな論理素子としたもの．
神経科学の知見を取り入れ，計算効率を優先した極めて大胆な工学的単純化である．

→★私の卒論「条件反射の生体工学的解析」の
2.3 節「シナプス結合による条件反射の解析」の図２－１で、
「McCulloch-Pitts の神経回路網のモデル」を掲載している。
（参考） https://www.1968start.com/M/bio/olduniv/soturon.htm

・実際のニューロンの確率的で複雑な挙動や，樹状突起における高度な非線形演算は，
現代の AI ユニット（積和演算としきい値処理）ではほぼ完全に無視されているが，
「入力の統合と出力の決定」という最小限の枠組みが，大規模なネットワークにおいて
驚異的な汎用性を示したことは，知能の計算論的側面を考えるうえで示唆に富む．

→★1960 年代の第一次 AI ブームのときの「積和演算としきい値処理」では、
しきい値を越えた素子間の結合係数の増分を学習効果としていた。

アーキテクチャ：

・ニューロンのネットワーク構造として、福島邦彦が提唱した「Neocognitron」は，
D.Hubel ら（1962）による視覚一次野の「単純型細胞」,「複雑型細胞」という
階層情報処理モデルから着想し，この階層を基本ブロックとして積み重ねることで，
画像認識を実現し、現代の畳込みニューラルネットワーク CNN の先駆けとなった．

学習則：

・D.Hebb[1949]が提唱した「Fire together, wire together」という原理は，
シナプスの結合強度が記憶や学習の基盤という現代神経科学のドグマを決定付けた．

→★卒論の参考文献

[15] D.O.Hebb: A textbook of psychology, W.B.Saunders company,1958.
（白井ほか訳、行動学入門、紀伊国屋）

（参考）https://www.1968start.com/M/bio/olduniv/soturon.htm

・現代の AI を支える誤差逆伝播法は，誤差信号を逆向きに伝える仕組みや重みの対称性の
要求から生物学的妥当性が疑問視されてきたが、近年，脳がバックプロパゲーションと
同等の数学的最適化を実現との可能性が議論されている．

→★下記の拙著（人工知能学会論文）の参考文献［14］で引用

Computational Semantics of a Neural Network System for Thought Process Simulation and its Applications, Journal of Japanese Society for Artificial Intelligence, 5, 5, 548-557 (Sep. 1990)

（参考文献 14）

Rumelhart,D.E.,Hinton,G.E. and Williams,R.J.: Learning Representations by Back-propagating Errors, Nature, Vol.323, pp.533-536 (1986).

（参考）https://www.1968start.com/M/p2/9009chuAI.pdf

・ニューロモジュレータやグリア細胞を介した大域的な信号が，シナプス結合の最適化に
重要な役割を果たしているとか，シナプス以外の記憶メカニズムの存在も議論され，
脳の学習は，多様な時間・空間スケールのメカニズムが重畳した複雑なプロセスである
と考えたほうが自然である．

→★もともと工学的応用として発展したニューラルネットワークの技術を
人間の脳の機能・構造の解明に用いるアプローチには違和感あり。

Attention：

・心理学の用語「Attention」が AI に借用される際，その機能的実態が変容している．
心理学における注意（attention）は，限られた情報処理資源を重要な対象に集中させる

選択的なボトルネック・メカニズムを指すが，AI における実装は，
入力データ間の関連性を重み付けして統合する動的な情報ルーティングに近い．

→★この指摘に関連するかわからないが、私の思考過程の計算モデルは、以下の通り：
＜拙著から引用＞
『思考言語に注目し、それを意識の面からとらえ、大脳におけるエネルギー分布の集中化作用とし、他方、集中したエネルギーの拡散化作用を連想機能として、これら両作用の交互反復過程を思考過程と考えた』

（拙著） **思考過程のシミュレ-ション、**
電子通信学会オ-トマトン研究会資料、A70-76(Dec.1970)
https://www.1968start.com/M/bio/olduniv/gakkai7012.html

・DNN は、脳そのものではなく研究者による仮説・モデル・比喩を出発点に，
工学的要請に沿った大胆な単純化と性能重視の設計によりが発達してきた．

・脳と AI の関係をより深く理解するためには，抽象的なモデルや概念化にとどまらず，
実際の脳活動データと AI モデル内部表現との間で，定量的かつ直接的な対応関係を
測定・検証する新たな研究パラダイムが必要となる．

→★私の修論の場合、「工学的要請」ではなく、「脳そのもの」すなわち人間の思考能力に興味があったので、討論学習や人間の思考能力の発達過程をシミュレーションした。

4. NeuroAI

・AI は脳に「インスパイア」されながら，工学的アプローチによって進化を遂げたが，
AI の内部構造に，現実の脳活動とのアラインメントが観察されるようになった．
AI と神経科学は新たな接点を迎え，「NeuroAI」と呼ばれる学際領域を生んだ．

・NeuroAI の源流となった初期の研究では，画像分類課題で訓練された DNN と，
同じ刺激に対する脳活動との関係を，三つの視点から検討した．

・第一は「神経エンコードモデル」．
DNN 内部の各階層の活動値から脳活動を予測する機械学習モデルを構築し，
脳のどの領野が DNN のどの層と対応するかを明らかにする．
D.L.K.Yamins ら[2014]は，サル視覚野の V4 は DNN 中間層，IT は DNN 最終層の活動で
最もよく予測されることを示し，脳の情報処理階層と DNN 層構造の対応関係を発見．

→★情報処理を【入力層｜中間層｜最終層】とみなせば、対応関係は当たり前では。

・第二は「表現類似性解析（RSA）」.
刺激セットに対する脳・DNN 活動パターンの表現構造を比較し，その相関で表現の類似性を評価．S.-M.Khaligh-Razavi ら[2014]は，教師あり学習で訓練された CNN が，IT 野の表現構造を最もよく説明することや脳と DNN の階層間対応関係を見いだした.

・第三は「神経デコードモデル」.
脳活動から DNN ユニット活動を予測するアプローチ．T.Horikawa ら[2017]は，
脳活動から層ごとの DNN 活動をデコードできることを示し，
階層が上がると解読可能な脳部位が高次領野へシフトすることも見いだした.

・神経科学の観点からは，NeuroAI の核心は，高次元の刺激・タスク空間で訓練された DNN の内部に現れる情報表現と脳信号の関係を直接・定量的に探ることにある.
訓練済み DNN を「実験動物」の脳のように扱い，その内部表現を分析することで，
DNN と脳のどこがどのように対応するかを検証する.

→★工学的観点で成果があれば、脳との関連付けは必要ないと思うが、
逆に脳の研究に役立つということ？　ちょっと違和感あり。

5. NeuroAI の理論的基盤

・脳と DNN の表現のアラインメントに対して，複数の理論的枠組みが提唱されている.

・脳の「メカニズム的説明」の妥当性を評価する基準として，D.M.Kaplan ら[2011]が提唱した「3M 基準（Model-Mechanism-Mapping constraint）」では，モデル内の変数のメカニズムの構成要素への対応と変数間の依存関係の因果関係への対応を要求する.
しかし，DNN の個々のユニットは生物学的ニューロンと 1 対 1 に対応していないため，厳密な 3M 基準では DNN は不適格と判断されがちであった.

・R.Cao ら[2024]の「3M＋＋フレームワーク」は，この制約を緩和する.
第一に，モデルが生の入力から出力を生成でき，脳活動を定量的に予測できれば，生物学的詳細をすべて再現する必要はない.
第二に，個体間の活動パターンが線形変換で関連付けられるように，
モデルから脳へのマッピングにも同じ変換クラスを適用する.
これにより，1 対 1 対応がなくても集団レベルでの機能的類似性を主張できる.
この枠組みにより，CNN などがサルやヒトの視覚野の説明モデルとして位置付けられる理論的根拠が提供されている.

・なぜアラインメントが生じるのかについては，「反変原理」は次のように説明.
単純なタスク（例：単純な幾何学図形の識別）に対する解決策は無数にあるが，

現実世界での物体認識のような困難で生態学的に妥当なタスクを解決できるシステムの構成方法は限られているので，タスクが十分に困難であれば，それを解決できるシステム（脳や AI）は，機能的に類似した特性をもつように「強制」される．
これは生物学における収束進化（例：魚とイルカの流線型の体）と似ている．

→★この説明には違和感あり。「魚とイルカの流線型の体」は適切な例ではないのでは？
タコやイカは例外？　鳥とジェット機やヘリコプターは反例になる？

・C. Conwell ら[2024]は，CNN と Transformer という質的に異なるアーキテクチャでも，脳活動の予測能力にほとんど差がなく，同じような表現形式に収束することを示した．
さらに，モデルが脳に似るかどうかの最大の要因の一つが学習データの多様性であり，多様な自然画像で訓練されたモデルが高い予測能力を示す．これは，
AI と脳がアラインメントを示すために，共通の経験が必要であることを示唆している．

→★AI と脳の類似性を示すために、まずは同じ学習データを用いるのは当たり前では？

・S. Onoo ら[2025]は，ニューラルな表現を、その因果的な起源ではなく，
その表現を「消費」する側が情報をいかに利用可能かという観点から再定義した．
通常，表現のアラインメントは，活動パターンの類似度（相関）で評価される．
これに対し，Onoo らが提唱する「読出し表現」は ，入力情報の再構成が可能な潜在空間内の点の集合として定義している．

・活動パターンの類似度が低くとも，読出し側が情報を高い精度で復元できるならば，
そこには豊かな表現が存在すると考える．実際，多様なモデルにおいて，
因果的な活動値から大きく逸脱した表現であっても，適切に読み出すことで
入力情報を極めて忠実度に再構成できることが定量的に示されている．

・これは，AI や脳による内部モデルが，単なる入力の抽象化や圧縮にとどまらず，
高次元の潜在表現空間に入力情報を頑健に保持していることを示唆している．
さらに注目すべきは，この読出しの枠組みにより，外部入力とは切り離されて機能する主観的イメージなどの心的表象も，外部刺激の有無にかかわらず潜在空間から
情報を「読み出せる状態」として，統一的に記述・議論できるようになる点である．

→★このあたりの話は門外漢にはよくわからない。(^^;;

6. 潜在表現の科学

・前章の理論的枠組みは，脳と AI の潜在表現がどのような意味で類似し得るのかを示す．

・本章では，著者の研究を中心に、脳と AI の潜在表現を計測・比較・応用する方法論を概観する．神経科学，コンピュータビジョン，自然言語処理が，表現類似性解析やエンコード／デコード，埋込み空間解析，線形読出しといった共通の手法や概念を共有するようになったことが，NeuroAI を支える方法論的基盤となっている．

・M. Schrimpf ら[2018]はモデルがどれほど脳らしいかを順位付けする枠組みにより、脳と AI の対応関係の体系的な評価を試みており、
物体認識のタスク精度が高いモデルほど脳との相関も高い傾向にあるものの，
最高性能の域ではその相関が飽和する現象が確認されている．

・同様に，S. Nonaka ら[2021]は脳領野とモデル層の階層的対応を評価し，
「高性能な DNN モデルは階層的には脳とあまり似ていない」ことを体系的に示した．
R. Geirhos ら[2018]の「CNN はテクスチャに依存し，人間は形状を重視」との指摘も，
入出力の正解率が等しくとも内部表現が本質的に異なり得ることを示している．
この比較は近年，大規模言語モデル（LLM）へと拡張され，視覚刺激に対する脳活動が，
その内容を記述するテキストの LLM 潜在表現と整合するといった，
モダリティを超えた表現のアラインメントも報告[2025]されている．

・読出し表現の観点では，潜在表現は単なる入力依存の活動値ではなく，
潜在空間から情報を再構成し得る機能的状態として捉え直される．この視点は，
脳からの視覚像のデコーディングや再構成に新たな位置付けを与える[2005, 2008]．

・K. Shirakawa らは，脳活動から DNN への予測を「脳から機械への潜在表現への翻訳」と再定義し，翻訳―生成パイプラインを提案[2025]した．この汎用的な枠組みにより，
知覚像だけでなく，想起像や錯視，注意イメージ，音声といった多様なモダリティへの拡張を統一的に扱うことが可能となる[2019～2025]．

・さらに，DNN の潜在空間を媒介とした脳コード変換技術は，個体や計測施設，
刺激条件の壁を越えたデコード・再構成モデルの広範な汎化を実現しており，
NeuroAI 研究の新たな共通プラットフォームとしての可能性を示した[2025]（図 1）．

図 1　脳・AI・心を横断する「潜在表現の科学」．

脳からの視覚像再構成を例に，脳と AI の潜在表現を介した情報の流れを示した．
脳から AI への潜在表現の「翻訳」により，主観的な視覚体験の客観的な外在化が可能．
下段の画像で，左列は被験者が実際に見た刺激，右列は fMRI 脳活動からの再構成の実例。
上段の画像は，仮想的な心の状態を表す．自然画像の知覚だけでなく，
物理的には存在しない主観的輪郭が知覚される錯視像も鮮明に画像として可視化される．
文献[Shen 2019, Cheng 2023, Kamitani 2025] をもとに作成。

・読出し表現と翻訳―生成パイプラインは，私的事象とみなされてきた主観体験を，
公的に検討可能な対象へと変換する方法論を提供．潜在表現空間の構造解析を通じて，
内観や言語報告に依存しない心理測定が現実味を帯び，脳・AI・心を横断する
統合的科学としての「潜在表現の科学」の可能性が具体化している[Kamitani 2025]．

→★ほんと？！

7. NeuroAI の課題

・NeuroAI は脳と AI の対応を定量化する強力な枠組みを提供してきたが，本章では，
予測精度，評価尺度，メカニズム，目的という四つの観点から現在の限界を整理する．

・第一に，予測精度の限界とノイズの解釈である．
神経エンコードモデルの「説明可能な分散」は 20%以下にとどまることが多い．

・第二に，ベンチマークと評価尺度の不完全性である．
現状のベンチマークデータセットは刺激の多様性に乏しい．

・第三に，理論的枠組みが，どこまで「メカニズムの説明」に迫れるかという問題．
DNN が前提とする「膜電位と化学シナプスによる情報伝達」という抽象化自体が，
実際の脳の演算プロセスを過度に省略している可能性がある．

→★私のモデルは、シナプス結合をニューロン間の結合係数としているので、
「実際の脳の演算プロセスを過度に省略している」指定はその通り。

・最後に，脳と AI の目的の差異と「帰納バイアス」の重要性である．
外界のモデル化を主眼とする AI に対し，脳が進化的に継承した帰納バイアスや，
発生・発達の動的なプロセスを考慮しないで，AI を脳の真のモデルとみなすのは困難．

→★脳の神経細胞の知見に基づいた計算モデルとしてのニューラルネットワークに
工学的意味があったからといって、「AI を脳の真のモデルとみなす」発想自体を
理解できない。

8. NeuroAI と心理学

・認知神経科学は，「注意」，「記憶」，「ワーキングメモリ」といった解釈しやすい
「構成概念」に依存するアプローチをとり、fMRI などで脳活動を計測して，
どの脳部位が「光る」かを調べる「脳機能マッピング」を行う．

→★「記憶」は比較的扱いやすいと思うが、それでも実際の人間の記憶想起に関する
多種多様な現象を説明可能なモデルの作成は、現状では困難と思われる

・NeuroAI は，構成概念を介在させず，DNN 内部に自律的に生じる「名もなき特徴量」と脳活動を直接的に対応付ける．構成概念に基づくトップダウンの認知神経科学と，名もな特徴量に基づくボトムアップの NeuroAI は，互いに相補的な関係にある．

・大規模言語モデル（LLM）を被験者として心理学的な実験を施す「AI 心理学」は，認知理論を検証するための独立した知性体として扱う．

9. おわりに

・「脳と AI は似ているか」という問いを起点に，AI と心理学・神経科学の関係を，その歴史的展開から現代の NeuroAI に至るまで概観してきた．

・計算資源と大規模データに基づく工学的成功が，「理論心理学としての AI」の理想を，かつてない精度で実現しつつあるという逆説的な現状がある．

・複雑なタスクを統合的に扱う NeuroAI は，心理学と神経科学をつなぐ新たな共通言語を提供している．

→★NeuroAI が『心理学と神経科学をつなぐ新たな共通言語』には違和感がある。
人間の心理の多種多様な現象と神経科学のミクロな知見との関連付けはできても、
このマクロレベルとミクロレベルの間の相互メカニズムを解明できるとは思えない。

・NeuroAI の真の意義は，「脳とそれほど似ていないモデルが，脳と似た内部表現を示す」理由を批判的に精査するプロセスにある．実際の脳との差分を埋めていく作業こそが，脳・AI・心を横断する「潜在表現の科学」を成熟させるために重要である．

→★「批判的に精査するプロセス」という指摘は、直前の私のコメントに近いのでは？
この表現と「新たな共通言語を提供」という表現は異なる意見のように思えるけど？

以上