2C5-OS-22b-1in

全脳アーキテクチャの解明を足がかりとした汎用人工知能の実現可能性

Feasibility of the Artificial General Intelligence based on the Elucidation of the Whole Brain Architecture

一杉裕志 *1
Yuuji Ichisugi

*1産業技術総合研究所
National Institute of Advanced Industrial Science and Technology(AIST)

Elucidating the details of the architecture of human's whole brain can be a promising approach to artificial general intelligence. While artificial general intelligence must be a useful technology, it also contains potential risks. Suitable regulations are required to prevent those risks to be realized.

1. はじめに

少子高齢化や資源制約等による生産性の低下を補う手段の１つとして、機械による労働の自動化がある。そのためには、従来人間にしかできないと思われていたタスクもこなせる機械、すなわち**汎用人工知能**の実現が必要となる。

それに向けたアプローチの１つとして脳の模倣がある *1。脳は、**大脳皮質、大脳基底核、海馬、小脳**など様々な器官から構成されているが、知能に関係する主要な器官の計算論的モデルは、不完全ながらすでに出そろっている。例えば大脳基底核は強化学習を行っており、報酬期待値を最大化する意思決定に関与すると考えられている。主要な器官がお互いにどう連携して脳全体の機能を実現しているかを解明し、解明された**全脳アーキテクチャ**を参考にして、計算機上で脳の機能を再現するというアプローチは有望であろう。

脳の計算論的モデルの進展の中でも特に重要と思われるのが、「大脳皮質はベイジアンネットである」という仮説の登場である [1] *2。大脳皮質とは脳の表面にある厚さ 2mm 程度の薄い組織で、知能に最も深く関与している。大脳皮質は、解剖学的な違いや他の組織との接続の仕方の違いによって**領野**とよばれる約 50 個の領域に区分けされている。認識、意思決定、運動制御、思考、言語理解といった脳の様々な高次機能が、このたった 50 個程度の領野のネットワークで実現されている。大脳皮質ベイジアンネット仮説は、この約 50 個の領野の動作原理に対して統一的な説明を与えつつある。

大脳皮質はヒトの全脳アーキテクチャのかなめであり、その動作原理を説明する仮説が現れてきたことは、全脳アーキテクチャを足掛かりとした汎用人工知能の実現に向けた大きな前進である。

2. 全脳アーキテクチャの特徴

ヒトの脳全体のアーキテクチャの特徴はなんであろうか。

脳を構成する主要な器官の働きは、**計算論的神経科学**においてはそれぞれ異なる機械学習装置として理解されている [2]。全脳アーキテクチャの特徴は、異なる特質を持った複数の機械学習装置を巧妙に組み合わせているところにあると思われる。

汎用性という観点から重要な特徴は、汎用の意思決定装置といえる大脳基底核と、汎用の連想記憶装置といえる海馬の両方を備えているという点である。これらはコンピュータに例えれば CPU とメモリに対応し、この２つをうまく使うことで脳は多様なタスクをこなしているのかもしれない。

もう１つの大きな特徴は、制御プログラムを自律的に学習して獲得していくという点である。この機能は定性的にはコンピュータ上で強化学習と教師なし学習で実現可能である。ここでは汎用の教師なし学習装置といえる大脳皮質が大きな役割を果たしているだろう。

定性的には脳のアーキテクチャとコンピュータに似たところがあるとはいえ、現状のコンピュータは脳ほど高い性能を発揮できない。多くの人はこの理由を脳の中にある未知なる情報処理原理に求めるかもしれない。しかし筆者はこの、現代の生気論とも言える「未知原理仮説」を支持しない。計算論的神経科学の分野においてこれまで数々の成功を収めてきたモデルは、いずれも計算機上で素直に再現できる極めて普通の情報処理モデルばかりだからである *3。

筆者は、脳の性能が極めて高い理由は、次節で述べる事前知識の作り込みにあると考えている。

3. 実環境に関する事前知識

ノーフリーランチ定理 *4 が示唆するところによれば、機械学習アルゴリズムの性能を上げるためには、学習の対象となる問題領域に関する事前知識を可能な限りシステムに作り込まなければならない。生物の脳は長い時間をかけた進化によって、実環境に関する事前知識を獲得し、それを使って性能を上げていると考えられる *5。

Deep learning 研究の第一人者である Bengio は、多くの成功した機械学習アルゴリズムをもとに、幅広く適用が可能な事

連絡先: 一杉裕志、茨城県つくば市梅園１－１－１ 中央第２産業技術総合研究所、 y-ichisugi@aist.go.jp

*1 参考：「全脳アーキテクチャ解明に向けて」 https://staff.aist.go.jp/y-ichisugi/brain-archi/j-index.html

*2 参考：「脳とベイジアンネット」 https://staff.aist.go.jp/y-ichisugi/besom/j-index.html

*3 ただしデジタル計算機上での効率的な再現が難しいモデルとしては、カオスを用いた計算論的モデルなどがある。しかし現時点では筆者はカオスが汎用人工知能の実現に不可欠とは考えていない。

*4 参考：「ノーフリーランチ定理 - Wikipedia」 http://ja.wikipedia.org/wiki/ノーフリーランチ定理

*5 なお、生物が置かれる実環境は、変化と多様性に富んでいるとはいえ、無限に多様というわけではない。脳は実環境に特殊化された知能を持つシステムなのであって、任意の環境に適応可能な万能のシステムでは決してない点は強調しておきたい。

前知識を整理して 10 個の**汎用事前知識** (generic priors) として分類している [3]。これには、滑らかさ、スパース性、階層性などが含まれている。この汎用事前知識のリストは現時点では不完全かもしれないが、将来実現される汎用人工知能の核心部分になると筆者はとらえている。

神経科学的知見は、生物が獲得した汎用事前知識が何かを教えてくれる。筆者が、複数の大脳皮質モデルにヒントを得て開発中である BESOM と呼ぶ機械学習アルゴリズム [4] には、様々な汎用事前知識が作り込まれている。

神経科学的知見は、約 50 個の領野それぞれが扱う視覚、聴覚、言語、運動制御などの対象ごとの**領域事前知識** (domain specific priors) についても多くのことを教えてくれる。

高性能な汎用人工知能を実現するためは、これら汎用事前知識と領域事前知識の両方の詳細を明らかにしていくことが最も重要であり、そのためには神経科学的知見と、学習対象の性質に関する工学的考察の両方が不可欠であろう。

4. 予想される人工脳の特徴

4.1 自然脳と人工脳の違い

脳が機械論的に動いているという前提（現時点ではこの前提を信じない人も少なくないが）が成り立っている場合、**人工脳**の振る舞いは**自然脳**と比べて何が同じで何が違うだろうか。

原則として人工脳は自然脳の能力を引き継ぐはずだが、生物学的制約がないところが大きな違いとなる。

また、もう 1 つの大きな違いは存在目的の違いである。自然脳は生物が子孫を確実に残すために必要な器官の 1 つとして進化してきたのに対し、人工脳は人間の役に立つようにエンジニアが発展させていくはずである。

これらの違いが具体的にどのような形に現れるかについて、筆者による現時点での予想を以下の節に簡単にまとめる。

4.2 人工脳が自然脳から引き継ぐ性質

人工脳の**知識発見能力や問題解決能力**は人間と同程度となるだろう。人間の脳は無限に高い能力を持っているわけでは決してなく、人工脳もそれを引き継ぐだろう。

常識については、人間と同じ環境で人工脳を持ったロボットが育つならば、人間と同じように身につくことになるだろう。最初は「物と物をぶつけると音がする」といった身近な常識から始まって、社会生活を続けるうちにより抽象的な常識を身に着けていくだろう。

自由意思、自己認識、創造性は、もし人間がそれらを持っているとするならば、人工脳も同じ程度に持つはずである [*6]。

4.3 生物学的制約がないことに起因する人工脳の特徴

思考速度、記憶力は、コストとのトレードオフだけで決まる。十分に高価なハードウエアさえ用意すれば、思考速度も記憶力も人間をいくらでも上回ることができる。

より重要な違いは、**寿命**がなく、**自己改変や自己複製**が容易である、という点である。これらの特徴は深刻な危険性をもたらす要因であり、慎重な対策が必要となるだろう。

4.4 存在目的の違いに起因する人工脳の特徴

感情や欲求 [*7] は、生物にとっては、子孫を確実に残すという目的のための機構として作り込まれている。人工脳の場合はそれをそのまま模倣する必要は全くなく、人間に役立つように感情や欲求を設計することになる。例えば、想定されたタスクをうまくこなせるときに快情動を感じるように報酬系を設計することになるだろう。

人工脳を備えたロボットが転倒などで容易に壊れないようにするためには、物理的衝撃を不快刺激として扱うなどの報酬系への作り込みが必要だろう。ただし、ロボット自身の**自己保存欲求**につながるこのような作り込みは、必要以上に強くせず、人間への貢献を優先するような設計がなされるべきである。

4.5 脳と互換系を持つ汎用人工知能の利点

全脳アーキテクチャに基づいて作られた汎用人工知能（人工脳）は、必然的に脳と似た振る舞いをするので、そのこと自体が利点をもたらす可能性がある。例えば人間にとって扱いやすく作られた道具を、たやすく利用するようになるかもしれない。また、人間向けに作られた教育カリキュラムや教材を再利用して、汎用人工知能の効率的な教育が行えるかもしれない。

5. 汎用人工知能の安全対策

機械の危険性を減らす上で重要な概念として、**本質安全**がある [5]。事故の被害を減らすために、機械の運動エネルギーなどの本質的な危険要因を小さくしておく考え方である。この考えに従えば、市場に出回るロボットは、運動能力・学習能力・推論能力などを必要最小限に抑えたものになるべきである。

また、一般に技術の危険性には、偶発的な事故（故障や暴走）と人為的な悪用の 2 種類があると思われる。両方の危険性を減らすために、高い知能を持ったロボットは潜在的に危険物であり武器であるという認識のもとで、開発・製造・保持に対する相応の規制が行われるべきである。

6. まとめ

脳の模倣は、汎用人工知能への有望なアプローチである。人工脳が実現されれば、あらゆる労働の支援に用いることで、人類は限りなく豊かになるだろう。ただし、富の再配分が正しく行われ、かつ資源制約の問題が解決されているという前提が必要かと思われる。人工脳の社会への影響については、社会科学の専門家と技術者を交えた真剣な議論が望まれる。

参考文献

[1] 一杉裕志, 解説 : 大脳皮質とベイジアンネット、日本ロボット学会誌 Vol.29 No.5, pp.412–415, 2011.

[2] 銅谷賢治, 臨時別冊数理科学 SGC ライブラリ 60 「計算神経科学への招待」脳の学習機構の理解を目指して 2007 年 12 月号.

[3] Yoshua Bengio, Deep Learning of Representations: Looking Forward, Statistical Language and Speech Processing, Lecture Notes in Computer Science Volume 7978, pp.1–37, 2013.

[4] 一杉裕志、「大脳皮質のアルゴリズム BESOM Ver.2.0」産業技術総合研究所テクニカルレポート AIST11-J00009, Sep 2011.

[5] 向殿政男、安全確保における本質安全の役割について, 検査技術, Vol.18, No,6, pp.26-31, 日本工学出版、2013-6.

*6 自由意思は、物理法則からの自由という意味であれば自然脳も人工脳も持ち得ない。しかし外界からの自由と解釈すれば、問題なく人工的に実現可能であろう。

*7 参考 :「感情や欲求の正体」
https://staff.aist.go.jp/y-ichisugi/rapid-memo/emotion.html