科学哲学科学史特殊講義（京都大学）

# 9　現在のAI：哲学的考察

**鈴木貴之**

（東京大学大学院総合文化研究科）

tkykszk@g.ecc.u-tokyo.ac.jp

# 1　現状の評価

ここまでの流れ：

古典的AI（Good Old Fashioned AI, GOFAI）：

- 記号システム仮説
- アルゴリズムとしての知能
- 問題：
  - scalability
  - 文脈理解

現在のAI：

- 背景：初期のニューラルネットワーク、機械学習
- 複雑な関数としての知能
- 問題：
  - 大量の訓練データの必要性
  - 変則的な振る舞い

古典的AIに対する批判：

- 意味理解には記号接地問題の解決が必要。
- 知能にはフレーム問題の解決が必要。

## 謎：問題解決なしの進展

- 現在のAIは、これらの問題を解決することなしにさまざまな成果を挙げているのでは？
  - トランスフォーマなどを用いた機械翻訳は、記号接地問題を解決していない。
  - AlphaZeroは、フレーム問題を明示的に解決せずにチェスや将棋における文脈の多様性に対応可能。

- 可能性：
  - これらの問題はじつは解決が不可欠な問題ではなかった？
  - 解決なしに計算量で克服可能？
  - じつは深層ニューラルネットワークはこれらを解決している？
    - 文脈の多様性は、変数の複雑な相互作用にほかならない。

# Dreyfus and Dreyfus 1988

> 「ここでの問題は、設計者が、ネットのアーキテクチャーによって、あるいくつかの可能な一般化は決して起こらないと決定してしまっていることである。…現実世界の状況においては、人間の知能の大半は、コンテクストに対して適切な仕方で一般化することで成り立っている。そこで設計者が、適切な応答のクラスをあらかじめ定義し、そのクラスにネットを制限するとすれば、その時ネットが表しているものは、そのコンテクストに応じて設計者がネットに組み入れた知能なのであって、ネットが常識を持つということにはならないだろう。」（邦訳、p. 56）

汎用AIと課題特化型AI：

- AI研究の究極目標は、人間のような知能をもつ汎用AIの実現。
- しかし、現在存在するAIは、ほぼすべて課題特化型AI。

→どうすれば汎用AIが実現できるだろうか？

うまくできる課題とできない課題：

- 翻訳／日常会話
- 画像診断／治療方針の決定
- 運転／部屋の片付け

違いは何か？

- 利用可能なサンプル数？
- 複雑さ？
- open-endedness?

何が必要か？

- 転移学習やメタ学習？
- ニューラルネットワークと記号的AIとの統合？
- 身体？

生物知能と人工知能：

- 生物個体はつねに自律的なエージェント。環境に対して適切な行動をとることが可能。
- メカニズムが複雑化するにつれて、知能はより高度になり、汎用性が高まる。
- 実世界で働くことが大前提。
- 知覚と運動が不可欠な要素。
- 生物には、痛みなどの形で報酬がhard-wireされている。

ロボティクス的アプローチ：

- AI研究も、現実世界で行動する身体をもつロボットをおもな研究対象とすべきでは？
- 比較的単純な昆虫（あるいはそれ以下の）レベルのエージェントから出発し、徐々に複雑化することで人間のようなエージェントに到達できるのでは？

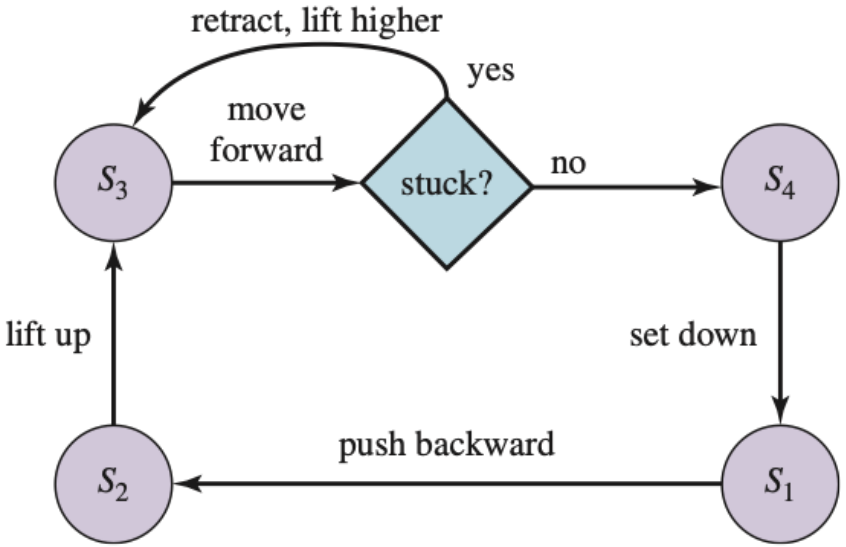
## Dreyfus and Dreyfus 1988

> 「ネットは、もしそれが、我々が持っている適切な一般化の感覚を共有するのなら、恐らくサイズ、アーキテクチャー、初期結合の状態を人間の脳と共有しているのでなければならない。また、もしそれが、すでに訓練者によって特定された連合をするように教えられるというのではなく、むしろ、それ自身の「経験」から、人間がするような連合をするように学習するのなら、ネットは我々が持っている出力の適切さの感覚をも共有していなければならない。

> そしてそれが意味するのは、ネットは我々の欲求、欲望、情動を共有していなければならないということであり、さらにまた、適切な肉体的運動、能力、そして、危害に対する可傷性などを備えた、人間が持っているような身体を、ネットが持っていなければならないということである。」（邦訳、pp．56–57）

# ブルックスのcreature

(a)




(b)

**Figure 26.32** (a) Genghis, a hexapod robot. (Image courtesy of Rodney A. Brooks.) (b) An augmented finite state machine (AFSM) that controls one leg. The AFSM reacts to sensor feedback: if a leg is stuck during the forward swinging phase, it will be lifted increasingly higher.

包摂アーキテクチャ：

- 単純な行動だけが可能な自己完結したエージェントを作成。
- 自己完結性を保ちつつ、行動を複雑化していく。
- 中央集権的な制御を行わない。局所的な制御の相互作用によって複雑な行動を実現する。
- 問題：包摂アーキテクチャは比較的単純なエージェントにのみ利用可能。

問題：

- 時間がかかる。
- このような手法でできるものは、生物と同様に、いろいろなことがそこそこうまくできるエージェント。これはそれほど有用ではない。
- 自律的なエージェントは、人間の希望通りに行動するとはかぎらない。これはそれほど有用ではない。
- 重要なステップ（記号的推論や言語仕様）をどのように実装すればよいのかは、いまのところ不明。

他の可能性：

- 人間とは別のアーキテクチャ？
  - 生物のアーキテクチャは進化の産物。知能を実現する唯一のアーキテクチャでも、最善のアーキテクチャでもないかもしれない。
  - 深層ニューラルネットワークは代替アーキテクチャ？
- 課題特化型の人工知能の統合？
  - 脳は深層ニューラルネットワークだが、機能分化が見られる。

# 2　主体としてのAIと道具としてのAI

主体としてのAIと道具としてのAI：

- Engelbartのintelligence amplification
- AI研究の2つのアプローチ（Copeland 1993, Ch. 2）
  - 人間の思考のシミュレーション
  - 人間の思考とは独立に有用なプログラム

- Hayes and Ford 1995:
  - 人間のような知性を作ることをAI研究の目標とすることは不適切。
  - 超人間的な知能を作るのではなく、人間の認知能力を拡張する人工物を作ることを目標とすべき。

人間の知能の欠点：

- 演繹的推論や確率的推論が苦手
- 認知バイアス
- 情動の影響
- ワーキングメモリの容量
- 疲労

→このような問題に対処するためにAIを活用するのが生産的では？

科学の道具としてのAI：

- 深層ニューラルネットワークなどによる高次元モデルを用いた科学の可能性。
  - 世界が単純な自然法則で説明できる保証はない。
- 説明と予測の分離？
  - メカニズムはわからないが結果は信頼できる道具をどこまで利用できるか？

# 3　心のモデルとしてのAI

深層学習は知能の基本原理か？

- 入出力モジュールのよいモデル？
- 問題：
  - 構造の違い
  - 学習メカニズムの違い
  - 必要な訓練サンプル数の違い

強化学習は知能の基本原理か？

- 神経科学的な妥当性（報酬系と予測誤差）
- 問題：
  - 複雑な計算
  - 必要な試行数の違い

そのほかの候補：

- ベイズ推論
- 予測誤差最小化

これらは相互に排他的なモデルではない。
→ニューラルネットワーク、強化学習、ベイズ推論の統合？

知能とは？

- 状況に対して（生存に十分な程度に）適切な行動を選択する能力。
- 人間が知能の典型と考える能力は、知能の本質ではないかもしれない。

- 脳は進化の産物。理想的なアーキテクチャを実現しているとはかぎらない。
- 知能に普遍的な特徴＋人間の知能に特有の特徴
  - 抽象度を上げれば、人間の知能とAIの知能に共通の特徴を見出すことが可能？
  - しかし、環境に対する適切な行動の生成、ベクトルの変換といった特徴づけは、抽象度が高すぎて有用ではないかもしれない。

- 世界のあり方＋エージェントの認知的資源→具体的な知能のあり方
  - 人間とAIの認知的資源に違いがあるとすれば、AIにおける知能の実現方法を人間も利用できるとはかぎらない。
- 問い：知能を実現する方法はただ一つか？