

人とロボットの相互理解を目指した心のアーキテクチャの提案

A Proposal of Architecture for Mind

Towards Mutual Understanding between Humans and Robots

小山 虎^{†‡}, 石黒 浩^{†‡}
Tora Koyama, Hiroshi Ishiguro

[†]大阪大学 基礎工学研究科

Osaka University, Graduate School of Engineering Science

[‡]JST ERATO 石黒共生ヒューマンロボットインタラクションプロジェクト

JST ERATO ISHIGURO Symbiotic Human-Robot Interaction Project

koyama@irl.sys.es.osaka-u.ac.jp

Abstract

In this paper, we propose an architecture for realization of the agent with the mind in an engineering way. This is an integration of the studies of the mind and self in philosophy and artificial intelligence, on the basis of their histories. We compare this architecture with existing architectures, such as reconstruction architecture and belief-desire-intention (BDI) architecture, and discuss the interdisciplinary significance of the architecture from the perspectives of robotics, philosophy, and cognitive science.

Keywords — Architecture for Mind, Robotics, Philosophy, Artificial Intelligence

1. はじめに

本論文では、心を備えたエージェントを工学的に実現するためのアーキテクチャを提案する。これは哲学と人工知能における心や自己の研究の流れを踏まえ、両者を統合したものである。

心は「人間とは何か」を明らかにする鍵であり、哲学では古くから研究対象となってきた。したがって、心を工学的に実現することができれば、「人間とは何か」の答えに大きく近づくことが期待できる。そのためには、心、そして自己のアーキテクチャが必要である。

心を工学的に実現するには、進化の歴史を再現するのが良いと考える研究者は少なくないだろう。心は人間が進化していく過程で獲得されたものであり、そのプロセスを再現できれば、最終的には心の獲得まで再現することが期待できるからである。しかし、このアプローチには深刻な問題がある。人間の進化は生命の誕生から数十億年をかけたプロセスである。一方、機械の誕生はせいぜい二千年前にすぎない。単純に進化のプロセスを再現するだけでは、数十億年とは言わないまでも、膨大な時間が必要になることが予想される。これを短縮するには、何らかのアーキテクチャを用いるのが有効であると考えられる。本論文では、哲学と

人工知能で研究されてきた心と自己のモデルを踏まえ、進化を直接の再現することなしに心を実現することが見込まれるアーキテクチャを提案する。

以下では、まず続く第2節で提案アーキテクチャの歴史的位置付けについて述べる。次に第3節でアーキテクチャの概要を説明し、第4節で既存アーキテクチャとの比較を行なう。第5節では提案アーキテクチャが進化の過程を踏まえたものになっていることを説明する。そして第6節では、提案アーキテクチャの特徴でもある二種類の欲求が必要な理由を、工学的観点と哲学的観点、そして認知科学的観点から議論する。最後の第7節では、全体のまとめと今後の課題について述べる。

2. 提案アーキテクチャの歴史的位置付け

前述のように提案アーキテクチャは、これまで哲学と人工知能で研究されてきた心と自己についてのモデルに基づいたものである。本節では、自己と心についての哲学と人工知能の歴史を概観し、提案アーキテクチャが歴史上のモデルとどのような関係にあるのかを説明する(図1)。

心と自己は人間の根本を形作るものであり、古代ギリシャの時代から現代に到るまで二千年以上も探求されてきた[1]。紀元前5世紀、哲学の創始者であるソクラテスは、「汝自身を知れ」という言葉を用いて自己理解の重要性を訴えた。ソクラテスの弟子であるプラトンはそれを踏まえ、人間の自己の本質は身体からは切り離される魂にあるという説を唱えた[2]。プラトンの弟子であるアリストテレスは、魂は身体から切り離すことができるようなものではなく、自己は魂と身体の両方から成る合成物だとしてプラトンに反対した[3]。

古代ギリシャの哲学者が考えていた自己は永遠不滅の魂という古い考えと結び付いていたが、17世紀フラ

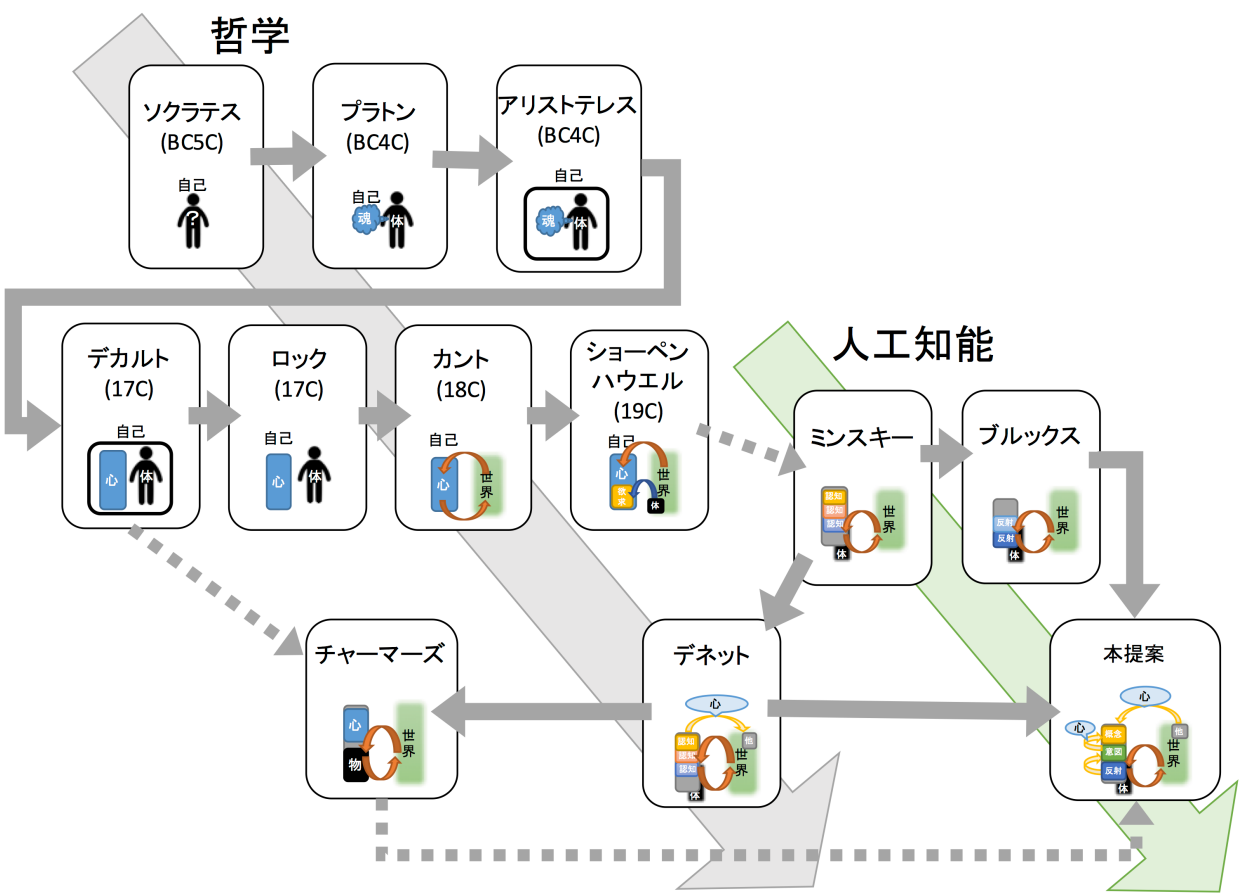


図 1 提案アーキテクチャの歴史的位置付け

ンスの哲学者デカルトがこれを現代的な考えに近づけた。デカルトは魂を心と同一視し、心は体と並ぶもうひとつの実体であり、人間は心と体という二種類の実体から構成されているという心身二元論を唱えた（その結果として、デカルトは心と体がどのような関係にあるのかという問題—心身問題—に直面することになった）[4]。イギリスの哲学者ロックは、体が入れ替わっても記憶や知識などの心的状態に変化がなければ自己は同じままであり続けると考えられることから、デカルトに反対し、自己にとって本質的なのは心だと主張した[5]。

デカルトやロックが考える心は理性、より具体的には知識や記憶と密接に関わっていた。18 世紀ドイツの哲学者カントはこれを発展させ、心の主な特徴は、(i) 外部世界から影響を受けて知識や記憶などの認識を形成すること、(ii) 認識から行動を引き起こすことという二つの能力にあるとした。すなわち、現代的な言い方をすれば、カントは心を認知的なものだと考えたのである[6]。これに対し、心には認知的ではない要素があると論じたのが 19 世紀ドイツの哲学者ショーペン

ハウエルである。ショーペンハウエルは大枠ではカントに従ったが、自己には、認知とは別に身体を突き動かす非理性的な「欲求」（ショーペンハウエル本人はこれを「意志」と呼ぶ）も必要であると考えた[7]。

19 世紀までの心は体とは独立に存在するものだと考えられていたため、心身問題は難問として哲学者を悩ませ続けた。それを大きく変えたのは 20 世紀以降の人工知能の発展である。なぜなら、人間と似たような人工知能を実現することができれば、心身問題の解決に大きく近づくことが期待できるからである。特に、ミンスキーの「心の社会」理論は、自己を認知エージェントとする点でカントやショーペンハウエルを引き継いでいるが、単なる理論に留まらない具体的なアーキテクチャを提案したという点で心と自己の研究にとって大きな転換をもたらした[8]。

人工知能における心と自己の研究が哲学に与えた影響の代表例はデネットである。デネットは、自己は人工知能で研究されているような認知エージェントであり、心は他者が自己の行動を社会的に解釈した結果だと主張する[9]。チャルメーズは、心を自己から引き離

すデネットに反対するが、人工知能研究と矛盾しないようにデカルト以来の二元論を復活させた[10]。

一方、人工知能研究では、実際に動作するロボットを作るという目的にとっては、外から得た様々な情報を複雑に処理して行動を決定する従来型の認知システムよりも、ブルックスのサブサンプションアーキテクチャに代表される、単純な行動から構成された行動ベースのシステムの方が明らかに有利であることが明らかとなり[11]、現在では行動ベースのアーキテクチャが支配的になっている。しかし、行動ベースのシステムでは人間のような高度な知能を実現することは困難であるため、心や自己のアーキテクチャとしては不十分なものに留まっている。

本論文で提案するアーキテクチャは、ブルックスのような行動ベースのアーキテクチャに、デネットのような行動を社会的に解釈した結果としての心を組み合わせることで、哲学における心と自己の研究と人工知能の研究の流れを統合したものになっている。

3. 提案アーキテクチャの概要

提案アーキテクチャの概要を図2に示す。本論文で提案するアーキテクチャは**行動・意図・概念**の三層構造をしており、それぞれのレイヤーは**操作**と**記憶**の二

つのモジュールから構成される。

3.1. 行動レイヤー

提案アーキテクチャの一番下にあるのは行動レイヤーである。行動レイヤーの操作モジュールは**反射行動ユニット**から成るネットワークであり、個々の行動ユニットは感覚情報を入力として運動機構を動作させる。また、行動ユニットには様々な重みがあらかじめ埋め込まれており、外部世界から得た情報を元にどの行動ユニットが選択されるかにはこの重みが大きく影響する。そのため、この重みは行動ユニットに本能的に埋め込まれた、ある種の欲求だとみなすことができる。

行動レイヤーの記憶モジュールは、人間の短期記憶に相当するものであり、感覚情報と実際に行なわれた行動が蓄積され、一定期間保存された後に破棄される。感覚行動記憶は単に貯蔵されるだけでなく、予測を行うために利用される。つまり、反射行動ネットワークは実際の感覚情報だけでなく、それと類似した記憶の後に記憶した情報も入力として処理する。

3.2. 意図レイヤー

行動レイヤーの上に位置するのは意図レイヤーである。意図レイヤーの操作モジュールは**複数の行動ユニ**

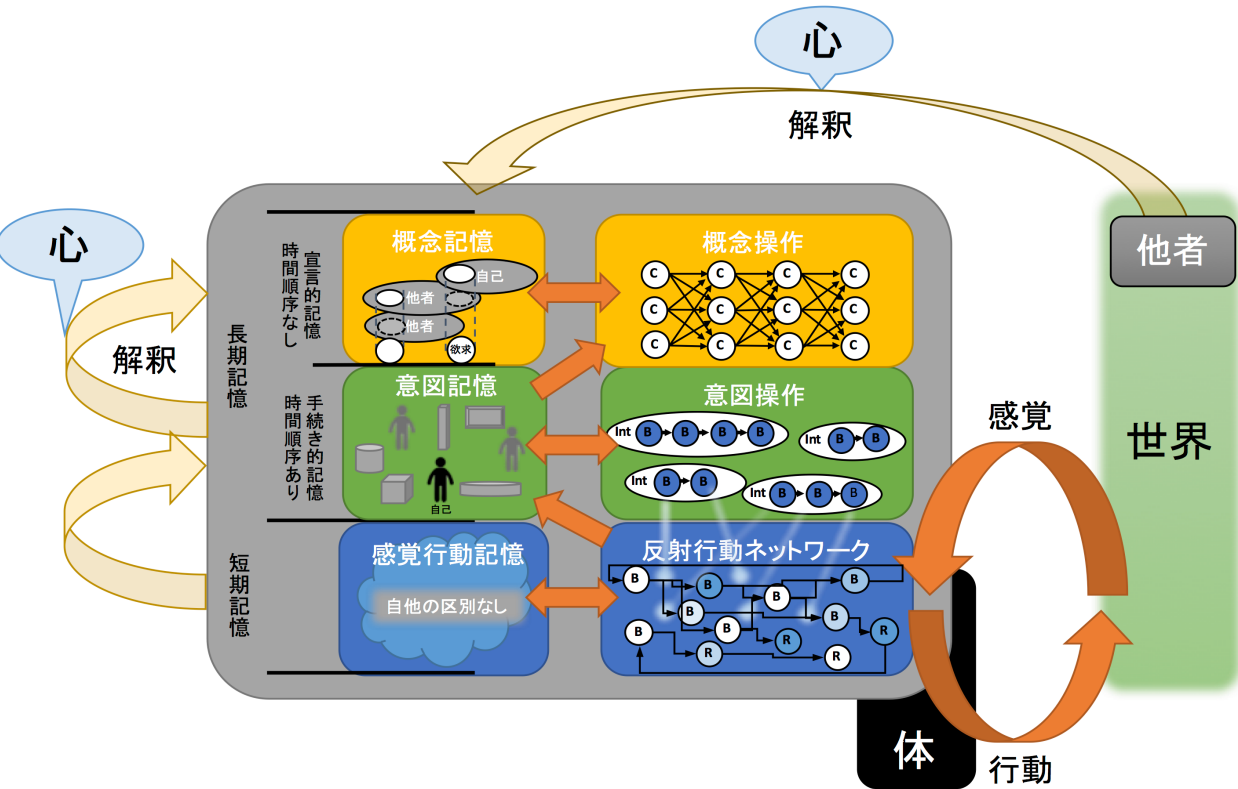


図 2 提案アーキテクチャの概要

ットを順番に繋いでひとまとめでした意図ユニットから成るネットワークである。アクティベートされた意図ユニットは、それを構成しているそれぞれの行動ユニットの実行されやすさを増加させる。したがって意図ユニットは行動計画とみなすことができる。

意図ユニットのアクティベーションには意図レイヤーの記憶モジュールが関わる。意図記憶は、自己とそれ以外の様々なオブジェクト、およびそれらの間のコンティンジェンシーと時間的順序関係から成る（限定的な）**世界のモデル**であり、反射行動ネットワークが行動感覚記憶を整理することで構成される。意図記憶は感覚行動記憶とは異なり、更新されるのみで破棄されない。この点で長期記憶のうちの手続き的記憶に相当する。

3.3. 概念レイヤー

一番上にあるのは概念レイヤーである。概念レイヤーの操作モジュールは**概念（記号表現）のネットワーク**である。個々の概念は言語コミュニケーションを介して獲得される。そのため、共起しやすい概念間の繋がりが強くなる。したがって、概念操作は一種の意味ネットワークとみなすことができる。

概念記憶は、概念操作が意図記憶に貯蔵されている情報から作り出す、**自己や他者など様々なエージェントから成る世界のモデル**である。概念記憶のエージェントは概念操作に含まれる概念の集まりとして表現される。このため、意図記憶のような時間順序はない。概念記憶は長期記憶のうちの宣言的記憶に相当する。

3.4. 心と概念記憶

提案アーキテクチャでは、心は複合的な概念である。すなわち、例えば、「ステーキが食べたい」という欲求（心的状態のひとつ）は、ステーキの概念と食欲（何かを食べたい）の概念が組み合わさったものである。欲求以外の心的状態も同様であり、例えば、「ライトを点灯しようとしている」という意図は、ライト点灯の概念と意図（何かを実現しようとする）の概念が組み合わさったものである。概念記憶は様々なエージェントから成るが、それらのエージェントを構成する概念にはこうした欲求や意図などの心的概念も含まれる。そのため、例えば、目の前にいる人に対応する概念記憶のエージェントは、ステーキが食べたいという欲求を含む概念の集まりとして表現されたりする。そう表現されるのは、意図記憶でその人に対応するオブジェ

クトの行動が当該の欲求に合致するときである。

意図や欲求を含む概念の集まりとして表現されるということは自己についても同様である。自己も概念記憶では一種のエージェントとして表現され、自己を構成するのは意図記憶での自己の行動と合致する概念である。前述のように反射行動ネットワークには本能的な欲求が埋め込まれているが、概念は言語コミュニケーションを介して獲得されるものであるため、本能的欲求と完璧に一致するような概念が獲得されることはほぼありえない。結果として、概念記憶での自己に含まれる欲求は、本能的欲求と完全には一致しないものの、本能的欲求に基づく行動によって獲得されるため、ある程度は本能的欲求を反映したものになる。

概念操作は、概念記憶での自己の欲求が満たされるように意図操作における意図ユニットのアクティベーションに影響を与える。これにより、本能的欲求や短いプランに従うだけでなく、より抽象的なレベルでの自己コントロールが可能になる。

4. 従来型アーキテクチャとの比較

本節では提案アーキテクチャを従来のアーキテクチャと比較する。比較するのは、(i)伝統的な再構成型アーキテクチャ、(ii)BDI アーキテクチャ、(iii)トップダウン型アーキテクチャの3つである（図3）。

4.1. 再構成型アーキテクチャ

伝統的なロボティクスでは再構成型アーキテクチャが採用されてきた。再構成型アーキテクチャでは、世界を完璧に再構成した内部モデルが最初に与えられており、感覚情報をこのモデルと照らし合わせ、定められた目的を達成するために必要な行動を記号操作によって計算する。行動には何も埋め込まれておらず、計算結果である行動が実行されるのみである。

再構成型アーキテクチャは環境が極めて限定されていれば十分有効である。しかし、完璧なモデルを再構成できていることを前提しているため、環境が多少なりとも複雑で、完璧なモデルをつくることができない場合に目的を達成することはかなり困難になるという問題がある。実社会のような極めて複雑な環境で動作するロボットのアーキテクチャとしてはあまり適切ではない。

4.2. BDI アーキテクチャ

近年の人工知能ではBDI (Belief-Desire-Intention)

アーキテクチャが採用されることも少なくない[12]。

BDI アーキテクチャの基本構成要素は信念(B)、欲求(D)、意図(I)という3つの心的状態である。信念は外部世界からの情報によって形成される外部世界の表現である(信念は誤っていることもありうる)。欲求は達成すべき状況を表現したものであり(ゆえにゴールとも呼ばれる)、一般的には少数の欲求があらかじめ与えられている。意図は欲求を実現させるために必要な一連の行動を表現したものであり、プランに相当する。BDI アーキテクチャでは、あらかじめ与えられた多くの意図(プラン)のうちのどれを実行するかが、信念と欲求を元に、BDI 論理と呼ばれる論理体系によって計算される。この行動決定プロセスは「熟考(deliberation)」と呼ばれる。

BDI アーキテクチャは人間の計画立案と行動決定について哲学者が考案した理論に基づいており[13]、適切な信念が形成されたときは適切な行動が選択される。また、誤った行動が選択される原因が理解しやすい(ほとんどの場合は信念に誤りがあることが原因である)。一方で、あくまで人間の合理的思考を模したものであり、行動ベースではない。そのため、実環境で動作するロボットに実装するには、行動ベースのアーキテクチャとうまく組み合わせる必要がある。

4.3. トップダウン型アーキテクチャ

ここでトップダウン型アーキテクチャと呼ぶものは、行動ベースのアーキテクチャの上に、意図と欲求のレイヤーを追加したものである。行動ベースのアーキテクチャは実環境で動作するロボットのアーキテクチャとして優れているが、そのままでは複雑な行動を実行することができない。トップダウン型アーキテクチャでは、追加された意図と欲求のレイヤーにより、欲求(ゴール)に基づくプラン実行が可能にする。

トップダウン型アーキテクチャでは複雑なプランの実行が可能になるが、適切な行動を選択するには信念(または部分的な世界のモデル)が必要である。また、それは欲求がオンラインになっている場合に限られ、オンラインになっていない場合は行動ベースのアーキテクチャと違いはない。したがって、トップダウン型アーキテクチャは、BDI モデルのような合理的エージェントと行動ベースのエージェントを切り替えるアーキテクチャだと考えてよい。

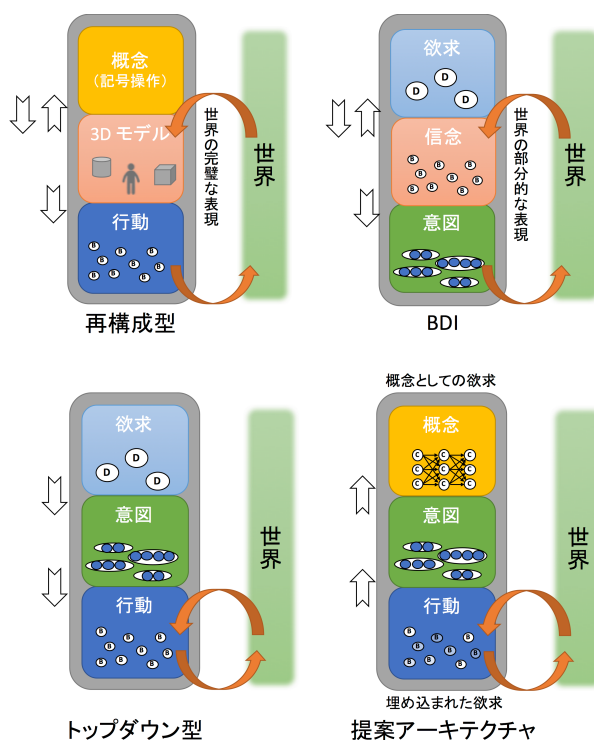


図3 四つのアーキテクチャの比較

4.4. 提案アーキテクチャとの違い

再構成型アーキテクチャと提案アーキテクチャの一番の違いはモデルである。再構成型アーキテクチャはモデルに基づいてしか行動決定ができないため、モデルが不可欠である。提案アーキテクチャもモデルを持つが、あくまで記憶を整理することで得られる限定されたモデルである。また、行動レイヤーは行動ベースのアーキテクチャであるため、行動決定にモデルは必須ではない。提案アーキテクチャのモデルは、複雑な行動を可能にするためのものである。

BDI アーキテクチャと提案アーキテクチャはどちらも意図と欲求を用いた行動決定をする点で共通している。また、提案アーキテクチャの概念にはBDI アーキテクチャにおける信念に対応するものも含まれていると考えられるため、提案アーキテクチャの概念レイヤーと意図レイヤーはBDI アーキテクチャの一種とみなすことも不可能ではない。

違いは、提案アーキテクチャが行動ベースのレイヤーを持ち、行動に欲求が埋め込まれていることである。まず、提案アーキテクチャには行動レイヤーがあるために熟考なしの行動決定が可能だが、BDI アーキテクチャには熟考が不可欠である。次に、通常のBDI アーキテクチャでは、欲求はあらかじめ与えられているが、提案アーキテクチャでは、欲求が埋め込まれているの

は行動であり、概念レイヤーでの欲求は行動を欲求という概念を用いて解釈した結果にすぎない。BDI アーキテクチャにおける行動は意図の構成要素にすぎないが、提案アーキテクチャにおける行動は、概念レイヤーにおける欲求がそれに基づいて解釈されるものであるという点で、このアーキテクチャの基盤になっている。このように提案アーキテクチャは、BDI アーキテクチャを包摂している。

5. 進化的妥当性：行動から概念へ

第1節で、進化の過程を再現して心を実現することは現実的でないのでアーキテクチャが必要だと述べた。しかし、提案アーキテクチャは進化をまったく無視しているわけではなく、むしろ進化的に見ても自然なものになっている。本節では、そのことを示すために、提案アーキテクチャが行動ベースのアーキテクチャからどのように発展したものかを説明する。

5.1. 行動と記憶

行動ベースのアーキテクチャとしてもっとも有名なのはブルックスのサブサンブションアーキテクチャである[11]。サブサンブションアーキテクチャは、それ以前の伝統的なロボティクスのアーキテクチャとは大きく異なり、反射行動を積み重ね、環境と一体化することで、メモリがなくても知的な行動ができることを示し、行動ベースのアーキテクチャの優位性を知らしめた。

しかし、サブサンブションアーキテクチャによって実現できた知性は限定的なものである。サブサンブションアーキテクチャを実装したロボットは、欲求が行動に埋め込まれており、適切な環境下で欲求に従った行動ができる。しかし、欲求に従って行動できることが知性のすべてではない。少なくとも人間が持っているような知性は、環境をモデル化し、手続きやルールを用いることで、複雑な行動を実現する。そのためには、モデル化に必要なメモリがなければならない。つまり、サブサンブションアーキテクチャに基づいて心を備えていると言えるエージェントを実現するには、少なくともメモリを備えている必要がある。

アーキテクチャにメモリを備えさせるにあたり、参照すべきなのは人間の記憶の研究である。実際、哲学における自己と心の研究においても記憶は重視されてきた（代表例はロックである[5]）。よく知られているように、人間の記憶は、まず短期記憶と長期記憶に分け

られる。短期記憶は短期間しか保持されないが、高解像度であり、感覚と密接に結びついていると考えられている。長期記憶は短期記憶とは違い、長期的に保持され、宣言的記憶と手続き的記憶に区別される。宣言的記憶は文章の形でデータとして保持され、手続き的記憶の方は、車の運転方法のようにスキルとして保存される。宣言的記憶はさらにエピソード記憶と意味記憶に区別される。エピソード記憶は自己の体験の記憶であり、時間や場所、感情と結びついて記憶される。それに対し、意味記憶は時間や場所、感情とは独立の抽象的なことがらの記憶である[14]。

こうした種類の記憶がどのように実現されているかは広く研究されているが、現時点では確実に正しいと言えることはまだそれほど多くはないようである。しかし、記憶がより高次の知的行動を可能にするために進化によって獲得されてきたものだとするれば、以下で説明するように、提案アーキテクチャに一定の進化的妥当性があることがわかる。

5.2. 記憶の獲得

サブサンブションアーキテクチャで実現されているような低次の知性を備えたエージェントは、行動に欲求が埋め込まれており、感覚情報に従って行動することで、欲求の充足によりもたらされる報酬が最大化されるように反射行動ネットワークを更新する。この種のエージェントにとって、最初に獲得しやすい記憶は短期記憶だと考えられる。短期記憶は感覚情報と密接に結びついており、また短期間で破棄されるため、一時的に情報を貯蔵するだけで実現できるからである。短期記憶の獲得には、外部世界から得られる感覚情報だけでなく、記憶に保持された情報もネットワークの更新に利用できるようになるため、行動の結果を予測した行動が実現できるようになるという利点がある。

長期記憶は短期記憶とは異なり複数の種類に区別されるため、段階的に獲得されたと考えるのが妥当である。最初に獲得されるのは短期記憶と同様に行動と結びついている手続き的記憶であろう。短期記憶と手続き的記憶の違いは、短期記憶では個々の行動が単位となる（それに伴う感覚情報が貯蔵される）のに対し、手続き的記憶では複数の行動のチャンクが単位となる点である。貯蔵する記憶の長期化にはこうした整理が不可欠だが、これにより、欲求と行動を一対一に対応させるだけでなく、特定の欲求を満たすために複数の行動をすることが可能になる。これが意図（プラン）

である。つまり、意図とは、記憶を利用した、欲求を満たすための短い計画行動だ、ということになる。

5.3. 概念の獲得と行動の予測

宣言的記憶は手続き的記憶とは異なる仕方では獲得されたと考えるべきであろう。なぜなら、宣言的記憶は概念に関わっていると考えられるからである。このことは、行動予測を向上させるために段階的に発展していったと考えることで説明できる。

短期記憶と手続き的記憶は行動予測に役立つ。何も記憶がなければ予測は不可能だが、短期記憶があれば記憶されている相手の行動が予測できる。また、手続き的記憶には、自己の計画行動（意図）の結果を予測できることが不可欠である（意図を選択できるのはその結果が予測できるからである）。しかし、短期記憶や手続き的記憶による行動予測には限界がある。どちらも自己の行動に基づく予測であるため、自己とは異なる行動が可能な他者の行動予測には失敗するからである。

他者の行動予測には概念を用いることが有効である。実際、意図や欲求といった概念は古代ギリシャの頃から、われわれ人間が行動予測をするために用いている道具立てとして知られている[15]。概念は言語コミュニケーションを介して獲得される。つまり、意図や欲求といった概念を獲得し、それを用いて他者をモデル化することは、他者による行動予測を言語コミュニケーションによって学習した結果だと考えられる。他者が実際に行っている行動予測の学習結果であるがゆえに、概念による他者の行動予測は有効なのである。

概念による行動予測は他者だけでなく自己にも適用可能である。自己の欲求や意図を直接知覚し、完全にモデル化することができれば自己の行動は容易に予測できる。しかし、これにはかなりのコストが必要となる。また、このモデルで行動が予測できるのは自己だけである。それに対し、他者の行動予測に有効な概念は、自己のモデル化にも用いることができる上に、すでに獲得された概念を利用するのだから、それほどコストはかからない。しかも、自己と他者の行動原理がまったく異なることがない限り（言語コミュニケーションが可能である以上、その見込みは低い）、他者の行動予測に有効であれば自己の行動予測にも有効であることが十分期待できる。

さらに、自己のモデル化により自己コントロールが実現できる。自己コントロールとは、欲求を認識しつ

つ、それを抑制する（その欲求に従った行動はとらない）ということである。行動ベースのアーキテクチャでは、たとえ記憶を備えていても、欲求に従わないことは別の欲求に従うことでしか実現できない。欲求の抑制には欲求以外の手段が必要なのである。概念によって自己の欲求をモデル化することにより、初めて欲求の抑制が可能になる。

6. ディスカッション

提案アーキテクチャの特徴は欲求が二種類あることである。また、この二種類の欲求は、通常は一致しない。そのため、提案アーキテクチャに従うエージェントには、自分が概念的に把握している欲求を満たすはずの行動をしても本能的欲求は満たされない、すなわち、自らの欲求に従おうとしても従えない、という事態が生じる。このようなパラドキシカルな状態が生じることは提案アーキテクチャの特徴である。本節では、二種類の欲求を措定する必要性をロボット工学、哲学、認知科学という三つの分野の観点から議論することで、提案アーキテクチャが特定の分野に留まらない学際的な価値を持っていることを示す。

6.1. 工学的観点からの議論

工学的観点からは二種類の欲求の必要性は、以下のように、①実時間性、②柔軟性、③共感性の三つの理由により説明できる（図4）。

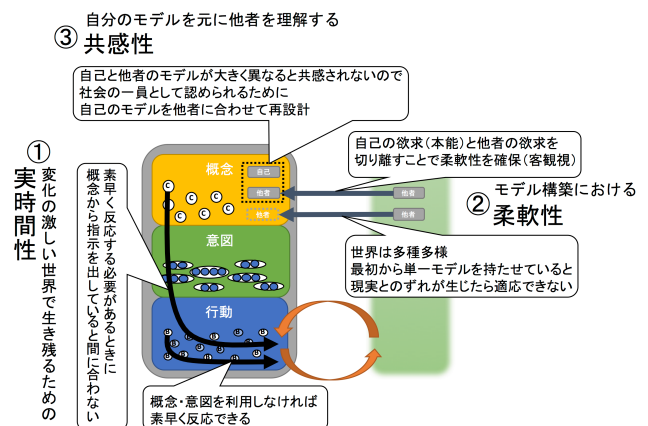


図4 工学的観点から見た二種類の欲求の必然性

① 実時間性

概念レイヤーは複雑な行動をするために用いられる。そのため、概念としての欲求を用いた行動決定は、行動レイヤーのみで行う（行動に埋め込まれた本能的欲求に従った）行動決定に比べると時間がかかる。常に

行動決定に時間をかけてよいならば、BDI アーキテクチャやトップダウン型アーキテクチャのように、常に欲求（概念）から行動決定を開始しても問題ないかもしれない。しかし、空腹時に食事をするときなど、行動決定に時間がかけられない状況も少なくない。そして、行動決定に時間をかけてよいかどうかの判定にも時間がかけられない（そもそも完璧に判定しようとするとはフレーム問題に陥る）。したがって、複雑な行動を実行するエージェントには、余裕があるときには概念から行動決定を開始するが、余裕がなくなれば自動的に行動レイヤーだけで行動を決定するようなアーキテクチャが必要になる。提案アーキテクチャでは、二種類の欲求により、これを実現している。

② 柔軟性

前述のように、再構成型アーキテクチャとは異なり、提案アーキテクチャでは世界のモデル化が完璧でなくても行動を決定できる。しかし、複雑な環境では、それに応じた複雑な行動をする必要があり、そのためには相応に複雑なモデルが必要になる。提案アーキテクチャのモデルは記憶に基づいて構成されるため、複雑なモデルをつくるには、実際の体験に基づく限定的な記憶を柔軟に活用しなければならない。このとき、概念としての欲求が本能的欲求を必ず一致する（すなわち、一種類の欲求しかない）のであれば、自己とは異なる欲求を持つ他者を含むモデルを作ることができなくなり、その結果、現実には確実に存在する自己と異なる欲求を持つ他者の行動に適応できなくなってしまう。本能的欲求と概念としての欲求を切り離し、後者の欲求を他者でも持ちうるものとして位置付けることにより、柔軟なモデル化が可能になる。さらに、自己が持ちうる欲求を他者が持ちうる欲求を区別しないことにより、自己の客観視が可能になるという利点もある。

③ 共感性

提案アーキテクチャは他者の存在、そして自己と他者を含む社会の存在を前提している。このことは、概念が言語コミュニケーションを介して獲得されるという点に反映されている。言語コミュニティ（社会）がなければ言語コミュニケーションも生じない。社会の一員として認められるには他者に共感されることが必要である。他者から共感されるために重要になるのは、他者と同じ行動原理に従っていることが他の者に伝

わることである。言い換えれば、他者から見て同じ欲求を持っているように行動することである。

そのためには、自己をモデル化する際に、純粹に自己についての記憶のみからモデル化するのではなく、他者に合わせてモデル化する必要がある。なぜなら、自己と他者の行動原理が共感を妨げるほど異なっている可能性があらかじめ排除されているわけではないため、自己についての記憶のみに基づくモデル化では、共感が得られない可能性が残るからである。概念としての欲求を本能的欲求から切り離せば、自己を他者に合わせて再設計することが可能なり、他者からの共感を導くように行動することも可能になる。

6.2. 哲学的観点からの議論

現代の哲学では、心については機能主義と呼ばれる立場が主流である[16]。機能主義によれば、心的状態はある種の機能的状態、すなわち、特定の入力に対し一定の出力を返す状態だとされる。機能的状態は計算論的に記述することが可能できるので実装可能であり、したがって人工知能研究と親和的だとされる。

機能主義にはいくつかの対抗理論があるが、代表的な対抗理論は解釈主義である（デネットも解釈主義者とみなされている[17]）。解釈主義によれば、心は他者の合理的に解釈した結果である。デネットは、他者の行動を予測する戦略として心的状態を帰属する「志向姿勢」があり、われわれ人間はその戦略を採用していると主張する。

機能主義と解釈主義は相容れない理論だとみなされている。しかし、提案アーキテクチャはこの両方を支持している。なぜなら、提案アーキテクチャは二種類の「心」を必要とするからである。例えば、行動に本能的に埋め込まれた欲求は機能的状態であり、機能主義を支持する。逆に、概念としての欲求は、まさにデネットの理論を念頭に置いたものであり、解釈主義を支持する。このことは、機能主義と解釈主義の論争が不適切であることを示唆する。提案アーキテクチャによれば、むしろ二種類の「心」があり、それが相互作用しているのが自己のあり方である。実際、二種類の心の相互作用は矛盾を生じさせるが、それが葛藤、迷い、意志の弱さなどの人間らしさを生んでいると考えられる[18]。

6.3. 認知科学的観点からの議論

認知科学的観点から見た提案アーキテクチャの意義

は「心の理論 (ToM)」に関わる[19]。ToM は他者の心的状態を表現する能力のことであり、ToM を持っているかどうかの決定的なテストは、他者に誤った信念を帰属する課題だと考えられてきた[20]。しかし近年、従来の言語を使った課題は不十分であり、言語獲得前の幼児も誤った信念を帰属することができるという実験結果が得られたため、言語に依存した ToM を「明示的 ToM」、言語に依存しない ToM を「暗黙的 ToM」として区別されるようになっている[21]。

明示的 ToM と暗黙的 ToM がどのように異なるのか、両者を基盤となっている能力は同一なのか、それとも異なるのか、という問いは、現在熱心に研究されており、まだ確定的な回答は得られていない。提案アーキテクチャは明示的 ToM と暗黙的 ToM がまさにまったく異なるものである可能性を示唆している。なぜなら、提案アーキテクチャでは、言語的な認知能力は概念レイヤーに関わるものであり、言語に依存しない能力はそれ以外のレイヤーで実現していることになるからである。提案アーキテクチャで生じるパラドックスは、言語的な認知能力とそれ以外の認知能力が区別されることで生じていると考えられる。

7. まとめと今後の課題

本論文では、心と自己についての新しいアーキテクチャを提案した。提案アーキテクチャは、二種類の相容れない欲求があるという点でパラドックスを生じさせるという特徴を持つ。また、提案アーキテクチャは哲学と人工知能における研究の流れを統合したものであり、再構成型アーキテクチャや BDI アーキテクチャなど既存のアーキテクチャが抱える難点を持たず、知的生物の進化の観点からも妥当なものである。

加えて、提案アーキテクチャは、その大きな特徴であるパラドックスが工学、哲学、認知科学という三つの分野の関連から必要性を議論することができるという点で、学際的にも大きな価値を持つものである。

残念ながら、提案アーキテクチャを実装したロボットの開発はまだ着手し始めたばかりであり、解決すべき技術的課題も少なくない。また、実装を進めていく中で本アーキテクチャの不十分な点も発見されると予想される。特に、レイヤー間の関係については、実装にあたってさらなる明確化と洗練が必要である。今後こうした改良を行なっていくことで、提案アーキテクチャの妥当性が示されると考えている。

参考文献

- [1] Sihvola, J., (2008) "Aristotle on the Individuality of Self," in P. Remes and J. Sihvola (eds.) *Ancient Philosophy of Self*, Springer.
- [2] Cooper, J., (1997) (ed.) *Plato: Complete Works*, Indianapolis: Hackett.
- [3] D. W. Hamlyn (1968) (trans.) *Aristotle: De Anima*, Oxford: Clarendon Press.
- [4] Descartes, R., (1641) "Meditation VI", in *Meditations on the First Philosophy*, J. Cottingham (trans.), Cambridge: Cambridge University Press, 1996.
- [5] Locke, J., (1689) *An Essay concerning Human Understanding*, P. H. Nidditch (ed.), Oxford: Clarendon Press, 1979.
- [6] Kant, I., (1781/1787) *Critique of Pure Reason*, P. Guyer and A. Wood (trans.), Cambridge and New York: Cambridge University Press, 1997.
- [7] Schopenhauer, A., (1818) *The World as Will and Representation*, J. Norman, A. Welchman, and C. Janaway (trans.), Cambridge: Cambridge University Press, 2010.
- [8] Minsky, M., (1986) *The Society of Mind*, Simon and Schuster, New York.
- [9] Dennett, D. C., (1987) *The Intentional Stance*, MIT Press.
- [10] Chalmers, D., J., (1996) *The Conscious Mind: In Search of a Fundamental Theory*, New York: Oxford University Press.
- [11] Brooks, R., (1989) "A robust layered control system for a mobile robot," in *IEEE Journal on Robotics and Automation*, vol. 2, pp. 14-23.
- [12] Singh, M., Rao, A. S., and Georgeff, M., (1999) "Formal Methods in DAI: Logic-Based Representation and Reasoning," In G. Weiß (ed.) *Multiagent Systems*, The MIT Press, pp. 331-376.
- [13] Bratman, M., (1987) *Intention, Plans, and Practical Reason*, Harvard University Press.
- [14] Squire, L. R., (2009) "Memory and Brain Systems: 1969-2009", *Journal of Neuroscience*, vol. 29(41), pp. 12711-12716.
- [15] Churchland, P. S., (1986) *Neurophilosophy: Toward a Unified Science of the Mind/Brain*. Cambridge: MIT Press.
- [16] Levin, J., (2016) "Functionalism", in E. N. Zalta (ed.) *The Stanford Encyclopedia of Philosophy* (Winter 2016 Edition), URL = <https://plato.stanford.edu/archives/win2016/entries/functionalism/>.
- [17] Child, W., (1996) *Causality, Interpretation, and the Mind*, Oxford University Press.
- [18] Stroud, S., (2014) "Weakness of Will" in E. N. Zalta (ed.) *The Stanford Encyclopedia of Philosophy* (Spring 2014 Edition), URL = <https://plato.stanford.edu/archives/spr2014/entries/weakness-will/>.
- [19] Premack, D. G., and Woodruff, G., (1978) "Does the chimpanzee have a theory of mind?", *Behavioral and Brain Sciences*, vol. 1(4), pp. 515-526.
- [20] Baron-Cohen, S., Leslie, A. M., and Frith, U., (1985) "Does the autistic child have a 'theory of mind'?", *Cognition*, vol. 21(1), pp. 37-46.
- [21] Low, J. and Perner, J., (2012) "Implicit and explicit theory of mind: State of the art", in *British Journal of Developmental Psychology*, vol. 30, pp 1-13.