汎用人工知能プラットフォームとしての人工意識

Artificial Consciousness as a Platform for Artificial General Intelligence

金井良太[1*]　藤澤逸平[1]　玉井信也[1]　眞方篤史[1]　安本雅啓[1]
Ryota Kanai[1]　Ippei Fujisawa [1]　Shinya Tamai [1]　Atsushi Magata [1]　Masahiro Yasumoto [1]

[1] 株式会社アラヤ
[1] Araya, Inc.

Abstract: In this paper, we propose a hypothesis that consciousness has evolved to serve as a platform for general intelligence. This idea stems from considerations of potential biological functions of consciousness. Here we define general intelligence as the ability to apply knowledge and models acquired from past experiences to generate solutions to novel problems. Based on this definition, we propose three possible ways to establish general intelligence under existing methodologies for constructing AI systems, namely solution by simulation, solution by combination and solution by generation. Then, we relate those solutions to putative functions of consciousness put forward, respectively, by the information generation theory, the global workspace theory, and a form of higher order theory where qualia are regarded as meta-representations. Based on these insights, We propose that consciousness integrates a group of specialized generative/forward models and forms a complex in which combinations of those models are flexibly formed and that qualia are meta-representations of first-order mappings which endow an agent with the ability to choose which maps to use to solve novel problems. These functions can be implemented as an "artificial consciousness". Such systems can generate policies based on a small number of trial and error for solving novel problems. Finally, we propose possible directions for future research into artificial consciousness and artificial general intelligence.

1　はじめに

映画などでは人工知能が自我に目覚め、ある一線を超える場面をよく見かける。この目覚めというのは、知的機能のの向上に伴って人工知能がある種の臨界に達したことにより、意識や意思を人工物が持つ瞬間を表している。映画などでは、意識をもった人工知能は、人間による統制からの自由を求めたり、人間的な感情にさいなまれたりするようになる。果たして、現在の人工知能技術の発展の延長で、このような意識を持った AI が生まれてくることには懐疑的な人も多いだろうが、そもそも、機能としての延長に意識があるという人間の直感には、どれほど妥当性があるのだろうか。このような問いに答えるためには、意識と知能の関係を理解する必要がある。漠然と、意識と知能には関係がありそうだと直感している人は多いのかもしれながい、意識が具体的にどのよな機能を持っているのかについては，現在のところ明確な答えはない。意識と知能の関係性については仮説すら十分に立てられていないのが現状である。

しかしながら、人工知能技術の発展が目覚ましい現在において、意識の機能を人工知能の観点から考え直す機運が高まってきている。人工知能の研究分野で利用される数理的な枠組みや物の見方は、意識の機能を考える上で新しい視点をもたらせてくれる。認知科学や神経科学における意識研究では、実は意識の機能が何であるかはあまり注目されてこなかった。というのも、意識の究極的な問題は主観的感覚（クオリア）がどのように、物理現象として閉じている脳のメカニズムから生まれてくるかというハードプロブレムこそが意識の本当の問題だと考える傾向が強かったからだ。しかし、今あらためて人

* 連絡先：株式会社アラヤ
東京都港区赤坂 1-12-32 アーク森ビル 24 階
E-mail: kanair@araya.org

工知能の研究で生まれてきている数理的な概念を利用しながら、意識の機能とは何かという問題に正面から取り組むことに格段の意義がある。意識の機能という文脈で、これまでの膨大な神経科学における意識研究の成果を再訪し、人工知能研究と対比させていくことで、意識と知能の関連性を明らかにすることができる。本稿では、意識の機能について仮説を立て、人工知能に汎用性をもたせるのに、どのような役割を果たしているのかを明らかにしていく現在進行中の議論の内容を報告する。

1.1 意識の科学

まず、意識の科学とはどのようなものだろうか。意識の問題とは、物質である脳からどのように「痛み」や「バイオリンの音色」といった感覚が生まれてくるのかという問いだが、これは「意識のハードプロブレム」と呼ばれ、現代の科学ではまったく太刀打ちできない難問だとも考えられている。というのも、意識というのは、主観的な存在であるため、客観的観測に基づく現在の物理学的世界観に収まらなそうに見える。特に、生物学などで成功を収めている、システムを要素に分解し下位の構造から上位の構造がどのように生まれてくるのかを明らかにする還元的アプローチは難しそうに見える。

しかしながら、意識という現象も自然現象であることには違いがない。つまり、意識もなんらかのこの宇宙を支配する普遍的な自然法則に従って発生しているはずだ。故に、意識の科学は、意識がどのような物理的または情報的条件を生じるのかを正しく特定するための自然法則の発見を必要としている [5]。このような意識の自然法則を仮定すると、その法則に従って条件を満たせば、地球上の生命の脳という器官以外でも、意識は生み出されるはずである。つまり、意識を生成するのに必要かつ十分な条件を満たす物理システムは、たとえ人工知能であっても内部経験としてクオリアを持つことが想定される[*1]。この観点からは、「人工意識」を作ることは原理的には可能である。しかし、ここで想定している意識の普遍的な自然法則がどのようなものであるかは、意識的経験が主観的にしか観測不能であるために、客観的観測を重視する現代的な科学の枠組みで取り組むことが困難となっている。

心の哲学においては、意識を機能的側面と主観的側面に概念上分離して考えることがある [4]。機能的側面はアクセス意識と呼ばれ、科学的精査の対象となる意識の客観的に観察可能な側面を指す。クオリアなどの意識の主観的な側面は、現象的意識と呼ばれ、その意識状態を経験している人を除いて直接観察することはできない。物理システム内で現象意識がどのように発生するかを理解することは、意識のハードプロブレムであり [6]、科学にとって最も難しい問題のひとつと考えられている。意識のハードプロブレムの可能な解決策として、往々にして意識自体には物理的な機能はなく、単に物理現象に付随している無機能な存在にすぎないという「随伴現象説」に帰着することがある。つまり、主観的経験は機能的な役割を果たさずに情報処理の副産物としてのみ存在するとされる [17, 27]。

この意識に機能はないのではないかという考え方は、最近の神経科学的研究により、多くの機能が無意識の内に処理が行われていることを示す実験証拠によりさらに強められている。例えば、注意は意識と非常に関連の深い機能だと考えられてきたが、注意は意識とは独立に機能する例が多数報告されてきている [3, 14, 19, 20]。さらに、ワーキングメモリ [26] なども無意識下におこるという報告がある[*2]。他にも指示に従って刺激に対する反応の仕方を制御するための実行制御機構も無意識の内に発動することを示唆する研究も発表されている [21][*3]。ここで上げたような意識研究の実験的成果もあり、意識を必要とする認知機能を限定するは非常に難しい。

そのような中で、いくつか意識を必要とするタスクというのは知られていた。その代表的なものは、古典的条件づけにおける「トレース条件づけ」と呼ばれる実験状況、それから、非反射的な行動を実行するためには、対象となる刺激の情報の意識的な保持が必要となることなどが知られている[*4]。

クオリアを最大の難問と設定している意識研究者は、意識の問題を単に意識の機能がどのように実現されているのかを解くだけの「イージープロブレム」をいくら解決しても意識の解明には至らないと考えるだろう。しかし、意識の機能の探求は、まだまだ未解決な問題で、情報が意識にのぼることで、どのような機能的利点が得られているのかを明らかにすることは、意識の理解のために重要である。この問題は、哲学者ダニエル・デネットにより「ハードクエスチョン」と名付けられており、ま

[*1] そのような立場は哲学においては生物学的自然主義とも呼ばれる [24]。

[*2] しかし、この無意識化でのワーキングメモリの効果は論文化されているデータでは非常に弱く、また、再現性を確認するための Pre-registration に基づいた実験では効果が認められなかったという報告もあるため、無意識化でのワーキングメモリの存在をどこまで鵜呑みにしてよいかは疑問がある。

[*3] このような実験も再現性があるのか検討を続ける必要はある

[*4] 視覚失認の患者 DF の研究では、視覚的に形状を認識できなくなってしまった患者でも、目の前にあるスリットの傾きに応じたアクションを行えるが、数秒の間、情報を保持した後にその形状の情報を利用して行動をするということができなくなった。

だまだ手つかずの状態で研究の進展が必要とされている。また、意識の機能を探求することは必ずしもクオリアを無視することではなく、クオリアについても機能という観点から、一歩踏み込むことができるのではないかと考えている。これについては後ほどメタ表現とクオリアに関する議論で触れる。

2　汎用知能を定義する

人工知能と意識の機能の関係を考える前段階として、汎用人工知能における「汎用性」とは何であるかを定義するところから始める。暫定的であれ、具体的な定義を与えることで、それがどのような方法で実現可能か議論することができる。汎用人工知能の構築に向けて、汎用性の定義がひとつの鍵となる。

汎用人工知能は人間レベルの理解能力と学習能力を備えたマシンであるという定義がある。しかしながら、人間をベンチマークとした場合には、人間の知能の特徴を抽出し定量化することが難しいため、汎用人工知能を評価するのに適していない。ここでは汎用性を次のように定義する。

定義 1. 知能の汎用性
過去の経験からの学習成果を利用して、新規のタスクを含む複数のタスクを効率よく解決する能力

このような定義では基本的には転移学習が主な機能ということになるが、このように定義しておけば汎用性は具体的に定量化可能となる。字義上も「汎用」あるいは General という概念とも親和性が高く、実現できたときの有用性も想像しやすい。このような汎用性の考え方は、Hutter が提案したフォーマルな汎用性の定義に関わる AIXI とも考え方は共通している [15, 16][*5]。人間における知能の汎用性を考えると明言し難いがより広義の機能を含むと考えられるが、ここでは敢えて汎用性を具体的に評価できるように狭義の定義を与えた。

3　汎用性を実現する方法

このように汎用性を狭く具体的に定義することで、今度はどのような解放 (Solution) があり得るのかを考えることができる。ここでは、汎用性を人工知能に実装するための方法を 3 つ提案する。それらは、決して網羅的ではないが、現在の技術の延長で実現可能であると考えられる方法を提案する。

汎用性実現方法 1. *Solution by Simulation*
フォワードモデル（あるいは世界モデル）を獲得していれば、報酬となる状態や目的が変化した場合でも、そこへたどり着くための方策を内部シミュレーションによって、臨機応変に学習することができる。この手法では、目的が変化しても、その目的にたどり着くための方策をその都度導出することができるため、過去に学習したモデルを利用することで将来の多様な課題に対する解決ができ，汎用性の定義を満たす。

このような手法で変化する目標に対して方策をアップデートするためには、前提条件として精度の高いフォワードモデルの獲得と未来の探索を効率よく行うためのメタ学習が必要となる。モデルベースの強化学習は基本的に Solution by Simulation の部類に入り、また近年の AI 研究においても数々の試みがなされている [13, 12]。このようなフォワードモデルの利用の仕方は、後に説明する意識の情報生成理論が想定している意識の機能と対応する。

汎用性実現方法 2. *Solution by Combination*
過去にに学習した特定のタスクに特化したモデルを複数用意しておき、それらを臨機応変に組み合わせてることで、新規の多様なタスクを、既存のモデルの組み合わせにより効率よく解決することができ、これは汎用性の定義を満たす。

より詳しく説明すると次のようになる。ニューラルネットワークというのは要は入力のベクトルを出力のベクトルへ変換する写像である。新しい課題を解くときに、個々の写像では対応できない変換が必要であったとしても、学習済みのネットワークの写像の合成によって解くことができる。例えば、画像 x からクラス分類 y を出力するニューラルネットワーク $f : x \to y$ と、クラス y を音声出力 z に移すネットワーク $g : y \to z$ を考える。f も g も、それぞれ特化した問題をネットワークではあるが、両者を合成した $g \circ f$ を構成することで、画像 x を入力として、そのクラス名を読み上げる合成ネットワークを新たに作ることができる。学習済みのネットワークを大量に持っていれば、この例のようにそれらを縦横無尽に組み合わせることで、多くの新しいネットワークを即時的に構成することが可能となる (図 1)。このような汎用性の実装方法の観点からは、タスクというのは学習済みネットワークによって繋がれた、有向グラフにおいて、任意のノード A からノード B への経路を見つけることに他ならない。ここでの提案方法と同一ではない

[*5] AIXI は強化学習における報酬を最大化するエージェントという観点から定義されており、現時点では他の形式化の可能性を残しておく。

が、これまで知られている研究成果で、学習済みのモデルの組み合わせにより新規の課題を解くというアプローチで最もよく知られたアーキテクチャとしては Pathnet があげられる [8]。

汎用性実現方法 3. *Solution by Generation*
ニューラルネットワーク自体を潜在空間に埋め込むことで、ニューラルネットワーク同士の関係性を表現した空間を作る。そのような潜在空間では各点が特定の機能を果たすニューラルネットワークに対応しする。その潜在空間からニューラルネットワークを生成することで、解くべき課題に応じて適切なニューラルネットワークを生成することができれば、既に学習したネットワークとの関係性より新規の問題に対して適切な解を出すネットワークを随時作ることができる。これは汎用性の定義を満たす。

これは蓄積した機能特化型のネットワークを元に、新規の課題を解決するという点では、Solution by Combination を更に拡張したものと捉えることができる。また、ここで提案しているようなネットワークを表現した空間を持つことは上記の Solution by Combination の課題を解決する場面でも役立つことは想定される。

Solution by Generation により、新規の課題に対して few-shot で識別機を直接生成してしまうようなメタ学習の手法はいくつか既に提案されている [22, 28, 11]。Meta Learning Autoencoder (MeLA) という手法 [28] では、分類問題を解くニューラルネットの重みを、潜在空間上の点を元に生成するネットワークと、データと正解データをこの潜在空間にマッピングするエンコーダを学習する。つまり、入力と出力の関係性をコンパクトに表現する潜在空間が構成されている。

4　汎用性と意識の機能の関係

前節では、汎用性の実現方法を 3 つ提案したが、これらは我々が想定している意識の機能と密接に関係している。その対応関係をこれから本節では論じていく（表 1）。

4.1　意識の情報生成理論

1 つ目の Solution by Simulation は、意識を必要とする認知課題を元に提案された「意識の情報生成理論」[18] と密接に関係している。意識の情報生成理論の主張は次のとおりである。

仮説 1. 意識の情報生成理論
意識の機能は、過去の環境とのインタラクションにより構築されたモデルを、現在の感覚入力とは切り離して、モデル自体の構造を利用して新たな情報を生成することである。そのようにモデルを利用することで、未来についてのシミュレーションに基づく行動プランニングや、現実に起きていない状況についての想像などが可能となる。

冒頭でも少し述べたが、意識と関連に深い機能として代表的なものとして、非反射的行動、トレース条件付け、短期記憶、行動のプランニングが上げられる。このような機能の共通項を分析することで、意識の本質的な機能は「反実仮想的な状況の感覚表現を内的なモデルに基いて生成する能力」であると仮説を立てたのが情報生成理論である。すなわち、現在目の前で起きていることではなく、数秒程度の過去や未来の出来事を、視聴覚等の感覚情報のフォーマットによって内的に生成することが意識の機能であると考えられる。そのような機能を実現する必要条件としては、感覚運動ループを通じた環境との相互作用により「自己」を含んだ生成モデルの獲得が必要である。

この「意識の情報生成理論」によって、意図・注意・思考といった主観的な心理状態に対してメカニズムとしての解釈を与えることができる。情報生成するエージェントとはすなわち内部でのシミュレーションができるので、新しい目標が設定されてもフレキシブルに対応ができ、新しい環境を効率的に探索する「好奇心」などの内発的動機を自然な形で実装できるという利点もある。

以上より、汎用性を内部の世界モデルを用いたシミュレーションにより獲得する Solution by Simulation は、「意識の情報生成理論」が意識の機能として位置づける、過去の経験から学習したモデルを外部から切り離して情報生成する機能と密接につながっていることがわかる。すなわち、「意識の情報生成理論」の観点からは、意識は汎用的知能を実現するための機能を提供していると言える。

4.2　グローバルワークスペースは学習済みモデルをつなぐ潜在空間

汎用性の獲得に、第 2 の手法である Solution by Combination を実現するためには、既存モデル同士での入力と出力のフォーマットが揃っている必要がある。様々な種類のデータを利用するためには、異なるモデルとの間での互換性が必要とされる。人間の脳によって実現されている意識においては、音声情報でも画像情報でも同一の主体が認識することができ、その意味では互換性が確立していると思われる。これまで、意識の機能として、

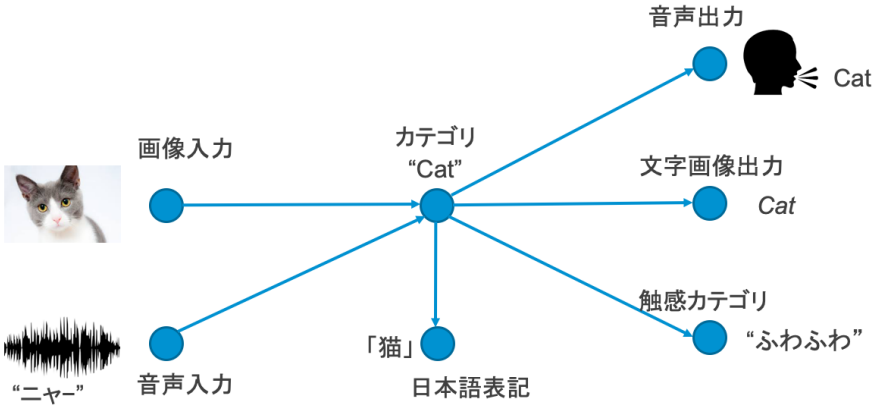

図1　組み合わせによる汎用性の獲得。各矢印は単一機能に特化した多層ニューラルネットを表している。汎用性はニューラルネットの有向グラフにおいて、任意のノード間での経路を確立することで実現される。

汎用性の実現方法	対応する意識の機能	意識の仮説
Solution by Simulation	反実仮想的シミュレーション	情報生成理論
Solution by Combination	共通の潜在空間を介した機能接続	グローバルワークスペース理論
Solution by Generation	メタ表現の空間を用いた機能の対象化	クオリアのメタ表現理論

表1　汎用性の実現方法と対応する機能と意識に関する仮説の対応について

脳内でのデータの互換性という観点ではあまり議論されてこなかったが、概念的には意識のグローバルワークスペース理論 [1, 2, 7] が関連が深い。グローバルワークスペース理論では、脳は特定の機能に特化したモジュールと、離れたモジュール間をつなぎ合わせる長距離結合に分けられる。特化型の単一機能では解決できない課題を解く際に、モジュール間での情報が共有されているグローバルワークスペースが必要となり、複数の特化型モジュールを協調させて課題を解決する。この際には意図的なエフォートが必要となる。また、特化型モジュールからの情報はこのグローバルワークスペースの中に入ることで意識にのぼると考えられている。

このような意識の機能の捉え方は、汎用性を実現する２つ目の方法である、Solution by Combination と合致するものである。しかしながら、具体的にブロードキャスティングとは何であるのかは明確にされてこなかったこともあり、また、特化型の機能モジュールを組み合わせて課題を解決することが、どのように実装されうるのかというところについても曖昧であった。しかしながら、汎用性を組み合わせで実現するという Solution by Combination という観点で、グローバルワークスペース理論を見てみると、具体的なイメージが浮かび上がってくる。しなわち、ブロードキャスティングというのは、特化型の機能モジュールの間での情報の連絡を可能とするために、モデル同士のデータ互換性を担保するメカニズムに他ならない。この仮説を明文化すると次のようになる。

仮説 2. 共有潜在空間としてのグローバルワークスペース

意識の機能は、多数の機能特化型のモデルの潜在空間を連結することでモデル間のデータの互換性をもたせることである。潜在空間の互換性をもたせることで、複数のモデルを自在に組み合わせて新しい機能を即時的に作ることが可能となる。脳内で意識の内容に貢献する部位としない部位の違い、すなわちグローバルワークスペースの内側と外側は、この潜在空間の共有化の範囲に含まれているかどうかで決まる。

このような異なる特化型モジュール（脳では視覚や聴覚などのモダリティ等）の間で潜在空間がシェアされていれば、転移学習などに利用することも容易となる。すなわち、意識とは多数の生成モデル群の潜在空間を統合したコンプレックスのことであり、この共有潜在空間が存在することで、モデル間の柔軟な組み合わせを可能となっている。

以上より、汎用性を既存のモデルの柔軟な組み合わせにより獲得する Solution by Combination は、「意識のグローバルワークスペース理論」が意識の機能として位

置づける、特化型モデルが単独では解決できない課題を組み合わせによって解決するプラットフォーム的な役割と合致する。すなわち、「意識のグローバルワークスペース理論」の観点からも、意識は汎用的知能を実現するための機能を提供していると言える。

4.3 ネットワークのメタ表現としてのクオリア

意識研究においては、意識が発生する条件について、大きく分けて 2 つの対立仮説がある。ここでは「一階表現理論 (first-order representation theory)」と「高階表現理論 (higher-order representation theory」と呼ぶ。一階表現理論というのは、意識的感覚すなわちクオリアが生じるためには、感覚情報を表現している状態だけで十分だという考え方である。つまり、視覚の主観的感覚というのは、視覚情報を処理することで得られる。しかし、一階表現理論は、なぜ脳の一部の処理は意識に上り、また別の処理（例えば、網膜上での視覚情報の変換など）は意識に上らないのかという疑問に対して答えがあまりない。高階表現理論というのは、それに対する答えとして、意識が生じるためには、単に感覚情報を直接表現しているだけでは不十分で、その表現に対するメタ表現が成立して初めて意識が生じると主張している。この高階理論には様々なバージョンがあるので、それらの差異について詳しくはここでは扱えないが、ここでのメタ表現というものが、「表現の表現」などと言われるものの、人工知能に実装しようとしたときに具体的にどのように定義されるものかが曖昧であった。すなわち、どのような数学的操作がメタ表現に対応するのかが明確ではなかった。

メタ表現を定義する試みとして次のような 2 つのケースが考えられる。表現というものを特に難しく考えずに、ニューラルネットによって入力 x を出力 y に移す写像 $f : x \to y$ を考え、y は x の表現だと考える。次に、もう 1 つのニューラルネットにより y を z に移す写像 $g : y \to z$ を考えれば、z は x の表現 y の表現になっているので、メタ表現になっているのではないか。しかし、これでは z も x を写像 $g \circ f$ で移した一階の表現に過ぎず、メタ表現と敢えて呼べるような質的な違いは起きていない。

認知神経科学分野での実験では、課題に対する反応が正解かどうかについての確信度 (confidence) を被験者に答えさせることでメタ認知を測るということが行われる。視覚刺激などの感覚信号に対しての弁別課題などで、無意識に行われた処理と意識に上った処理の間で、主観的な確信度に違いが現れるため、主観的な知覚が生じたのかどうかの判別にも用いられている手法である [25, 9, 23]。計算論的には、確信度は不確かさ (Uncertainty) の推定を意味する。これは、心理学的な実験において使われてきたという経緯はあるが、実装するという観点からは非常にシンプルなもので、ニューラルネットワークの文脈では、出力層の softmax を計算することで擬似的な確率として扱うことで、予測の確信度として頻繁に利用されている。往々にして、softmax で算出した確信度が高すぎるという問題はあるが、それを調整するような手法も提案あされている [10]。では、確信度を計算するという処理が追加されれば、メタ表現ができたと言えるのだろうか。

このような観点から、メタ表現について満足のいく数理的定式化をすることが難しかったが、前節で示した汎用性の実装方法である写像の埋め込みをメタ表現の実現方法と捉えると非常にスッキリする。

仮説 3. クオリアのメタ表現仮説
脳内には脳部位間の結合として存在するニューラルネットワークの入力と出力の関係をメタに埋め込んだ表象が存在し、この埋め込み空間（クオリア空間と呼ぶ）における各座標が、それぞれのネットワークのメタ表現となっている。我々が感覚の質（あるいはクオリア）と呼んでいるいるものは、この空間に表現されているネットワークのメタ表現のことである。

ニューラルネットが表現している入力と出力このメタ表現を持つことで、一階表現であるニューラルネットによる識別プロセスを別の空間における対象としてみなすことができる。そのようなメタな表象を持つことで、1 次のネットワーク間の類似性や構造が表現できる。すなわち、メタ表現を持つことで、赤と紫が主観的感覚において、似ているかどうかなどの評価が可能となる。この仮説の言わんとしていることは、次のようなことである。私たちが赤クオリアを感じるというのは、赤を抽出するフィルターを持つことで赤を見ているのではない。むしろ、赤を抽出するフィルターを潜在空間に埋め込むことで、赤の赤らしさというものを対象化している。そうすることで、この赤という情報処理が他の感覚とどのような関係にあるのかを、埋め込まれた潜在空間において比較することが可能となる。つまり、一階表現では異なるフィルター同士の関係は潜在的に存在しているだけだが、それらの関係を明示した空間を作ることで、その処理の質をクオリア空間に位置づけることができる。そのような表現を持つことのメリットは、前節における汎用性の実現手法における Solution by Generation（生成による解決）を実現できる点である。つまり、新しい課題

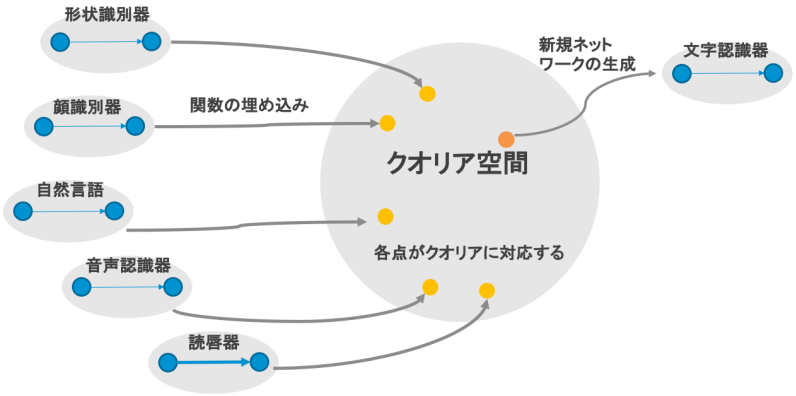

図2　クオリア空間。ニューラルネットワークの埋め込み空間が、一階表現のメタ表現を構築することで、ネットワークがもっている内在的な特性についての類似性や質に関する方向性を捉える空間をクオリア空間と呼ぶ。このような潜在空間をもとに、ネットワークの生成モデルを持つことで、任意の新規の機能について既に獲得しているニューラルネットワークの集合からの類推で新しい機能を生成することができる。このような枠組みは MeLa などのメタ学習の取り組みでは既に知られている。

が与えられたとき、どのネットワークを使えばよいのか、クオリア空間上で選択が可能となる。

5　意識と知能をつなぐ展望

本稿では、汎用性を「過去に学習した機能を利用して、新しい問題を効率よく解く能力」と定義し、一方で、意識の機能はモデルを入力とは切り離して自由に組み合わせて利用するプラットフォームとして機能すると提案した。その上で、意識は限られたリソースを組み合わせることで、より広い汎用性を実現するためのプラットフォームとして機能すると提案した。ここでの議論をもとに、実装はある程度可能であると想定している。表 1では、この論文で提案した 3 つの汎用人工知能の構築方法が、それぞれ、意識のどのような機能と対応しているのか、対応する仮説の名称とともに示している。

今回、 3 つの汎用性実現方法を提案し併記したが、相互の依存関係や共通性についてはさらなる議論と考察が必要である。例えば、Solution by Combination と Solution by Generation では、実現方法が違えど、どちらも既存の関数がもつ性質をメタ表現することが重要なエッセンスとなっていると考えられ本質的に両者は区別されるべきものなのかなどの疑問も湧いてくる。意識の理論についても、情報生成理論で述べているような、既存の学習済みモデルを現在の感覚入力だけではなく、架空の状況に対して再利用するというのは、グローバルワークスペースで複数のモデルを組み合わせて機能を実現するための必要なステップとなっており、包含関係にあるのではないかとも考えられる。このように、汎用性の実現方法についても、意識の理論についても仮説間での関係性を精査し、本質的に重要なメカニズムが何であるかをさらに整理していく必要があるだろう。

もう 1 つの重要な今後の展望は、ここで上げた汎用性の実現方法をニューラルネットワークにより実装し、適切な汎用性の評価課題のもとで、スケーラブルに新規の課題を解決することができるかを検証していくことである。ここで示してきた汎用性の実現方法は概念的なレベルでの記述にすぎない。実際に、どのような実験環境においてならば、有効性が示せ、どのような場合に困難が潜んでいるのか、実験を通した検証が必要である。

6　まとめ

本稿では、汎用人工知能を実装する方法を 3 つ提案し、それぞれが意識の生物学的機能とどのように結びついているかについて議論した。汎用性の実現という文脈の中で、意識の機能を考えることで、これまで意識の理論的研究において提案されてきた概念（例えば、グローバルワークスペース理論における「ブロードキャスティング」や、意識の高階表現理論における「メタ表現」）に対し、具体的な機能的意味を付与することができた。さらに、これまで意識と知能という関連性は明確に説明することが難しかったが、本稿では両者の具体的な関係性を提案に至った。このような試みは、詳細を欠いていた意識の

科学的理論の洗練化に貢献し、汎用人工知能の具体的な実装デザインとしての役割を果たす。このように、意識の機能を解き明かしハードクエスチョンを解いていくことで、「人工意識」の構築と、それによって実現される汎用人工知能開発の礎としたい。

謝辞

本研究の着想は、JST、CREST（JPMJCR15E2）の人工意識に関するプロジェクト及び NEDO の人工意識に関するプロジェクトについての議論の中で生まれた。

参考文献

[1] B. J. Baars. In the theatre of consciousness. global workspace theory, a rigorous scientific theory of consciousness. *Journal of Consciousness Studies*, 4(4):292–309, 1997.

[2] B. J. Baars. Global workspace theory of consciousness: toward a cognitive neuroscience of human experience. *Progress in brain research*, 150:45–53, 2005.

[3] B. Bahrami, N. Lavie, and G. Rees. Attentional load modulates responses of human primary visual cortex to invisible stimuli. *Current Biology*, 17(6):509–513, 2007.

[4] N. Block. On a confusion about a function of consciousness. *Behavioral and brain sciences*, 18(2):227–247, 1995.

[5] D. J. Chalmers. *Toward a theory of consciousness*. PhD thesis, Citeseer, 1993.

[6] D. J. Chalmers. *The conscious mind: In search of a fundamental theory*. Oxford university press, 1996.

[7] S. Dehaene, M. Kerszberg, and J.-P. Changeux. A neuronal model of a global workspace in effortful cognitive tasks. *Proceedings of the national Academy of Sciences*, 95(24):14529–14534, 1998.

[8] C. Fernando, D. Banarse, C. Blundell, Y. Zwols, D. Ha, A. A. Rusu, A. Pritzel, and D. Wierstra. Pathnet: Evolution channels gradient descent in super neural networks. *arXiv preprint arXiv:1701.08734*, 2017.

[9] S. M. Fleming and H. C. Lau. How to measure metacognition. *Frontiers in human neuroscience*, 8:443, 2014.

[10] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 1321–1330. JMLR. org, 2017.

[11] N. Guttenberg and R. Kanai. Learning to generate classifiers. *arXiv preprint arXiv:1803.11373*, 2018.

[12] N. Guttenberg, Y. Yu, and R. Kanai. Counterfactual control for free from generative models. *arXiv preprint arXiv:1702.06676*, 2017.

[13] D. Ha and J. Schmidhuber. World models. *arXiv preprint arXiv:1803.10122*, 2018.

[14] P.-J. Hsieh, J. T. Colas, and N. Kanwisher. Pop-out without awareness: Unseen feature singletons capture attention only when top-down attention is available. *Psychological science*, 22(9):1220–1226, 2011.

[15] M. Hutter. A theory of universal artificial intelligence based on algorithmic complexity. *arXiv preprint cs/0004001*, 2000.

[16] M. Hutter. *Universal artificial intelligence: Sequential decisions based on algorithmic probability*. Springer Science & Business Media, 2004.

[17] F. Jackson. *Epiphenomenal qualia*. Abr, 1982.

[18] R. Kanai, A. Chang, Y. Yu, I. M. de Abril, M. Biehl, and N. Guttenberg. Information generation as a functional basis of consciousness. *Neuroscience of Consciousness*, in press.

[19] R. Kanai, N. Tsuchiya, and F. A. Verstraten. The scope and limits of top-down attention in unconscious visual processing. *Current Biology*, 16(23):2332–2336, 2006.

[20] C. Koch and N. Tsuchiya. Attention and consciousness: two distinct brain processes. *Trends in cognitive sciences*, 11(1):16–22, 2007.

[21] H. C. Lau and R. E. Passingham. Unconscious activation of the cognitive control system in the human prefrontal cortex. *Journal of Neuroscience*, 27(21):5805–5811, 2007.

[22] A. A. Rusu, D. Rao, J. Sygnowski, O. Vinyals, R. Pascanu, S. Osindero, and R. Hadsell. Meta-learning with latent embedding optimization. *arXiv preprint arXiv:1807.05960*, 2018.

[23] K. Sandberg, B. Timmermans, M. Overgaard, and A. Cleeremans. Measuring consciousness: is one measure better than the other? *Conscious-*

ness and cognition, 19(4):1069–1078, 2010.

[24] J. R. Searle. Biological naturalism. *The Blackwell companion to consciousness*, pages 325–334, 2007.

[25] M. T. Sherman, A. B. Barrett, and R. Kanai. Inferences about consciousness using subjective reports of confidence. *Behavioral methods in consciousness research*, pages 87–106, 2015.

[26] D. Soto, T. Mäntylä, and J. Silvanto. Working memory without consciousness. *Current Biology*, 21(22):R912–R913, 2011.

[27] M. Velmans. Is human information processing conscious? *Behavioral and Brain Sciences*, 14(4):651–669, 1991.

[28] T. Wu, J. Peurifoy, I. L. Chuang, and M. Tegmark. Meta-learning autoencoders for few-shot prediction. *arXiv preprint arXiv:1807.09912*, 2018.