The 36th Annual Conference of the Japanese Society for Artificial Intelligence, 2022

意識の Entification 仮説

Exploring the Hypothesis that Consciousness is an Entification

山川　宏[*1, *2]　Hiroshi Yamakawa
布川　絢子[*2]　Ayako Fukawa
松尾　豊[*1]　Yutaka Matsuo

[*1] 東京大学　The University of Tokyo
[*2] 全脳アーキテクチャ・イニシアティブ　The Whole Brain Architecture Initiative

In current consciousness research, the major theories that originate in neuroscience deal with the integration of information, and many of the theoretical arguments originate from existence in the world. However, none of these approaches provide sufficient insights into computational mechanisms. In contrast, the entification process, in which the search for existence is performed on a data structure (aligned structure) that allows inductive reasoning, can provide a basic mechanism for information integration to recognize existence. In this study, we set up a working hypothesis that consciousness is an entification process. Next, we will consider the multifaceted position of the working hypothesis and discuss approaches to studying the brain as a reference for the mechanism that performs the entification process.

1. はじめに

最近、世界の中に存在を見出す処理である Entification の定式化が進められている[山川, 2022]。そこで Entification とは、世界についての予測性が高いコンパクトなモデル(存在)を含むように、その基盤となる整列構造の探索も同時に行う処理である。ここで整列構造とは、帰納推論を可能とするデータ構造である。

複雑な世界に対処するために世界モデルをもつ知能システムにとって、その中に存在に対応する個別のモデル(Entity 記述)を持つことで得られる利点は、Entity 記述を用いた汎化能力の向上に留まらない。まず世界に対する記述を、その構成要素となる Entity 記述を組み合わせて行うことで記述量や思考空間を削減できる。さらに、Entity 記述は、記号で指示可能な対象となるのでそれは言語の基盤となるなどの利点がある。

他方で、現在の意識研究では、理論的な議論の多くは世界の中の存在を扱い、神経科学に端を発する主要な理論では情報の統合を扱っている。しかし、いずれのアプローチも、計算論的な機構については十分な示唆が得られていない。

本研究では、次の 2 章で意識を Entification 処理とみなす「Entification 仮説」を作業仮説として設定し、次の 3 章では、その作業仮説の位置づけなどについて議論する、さらに 4 章において Entification 処理を行う機構に対して脳を参考として研究するアプローチについて検討を行い、最後にまとめる。

2. 意識の Entification 仮説

2.1 整列構造

帰納推論を実行する際は、関係モデル[Codd,2002]を基礎とする関係データベース(RDB)や表形式といったデータ構造を用いる。

RDB のデータ構造は図 1A に示すよう、項目毎の列にあたるカラムと、事例毎の行にあたるレコードと、レコード内の1つの要素にあたるフィールドからなる。各フィールド内に1つの値が保持される。各レコードを一意に特定できるカラムとして主キーがある。これら全体がテーブルと呼ばれる。レコードとカラムの何れも集合であり、行と列は順不同である。

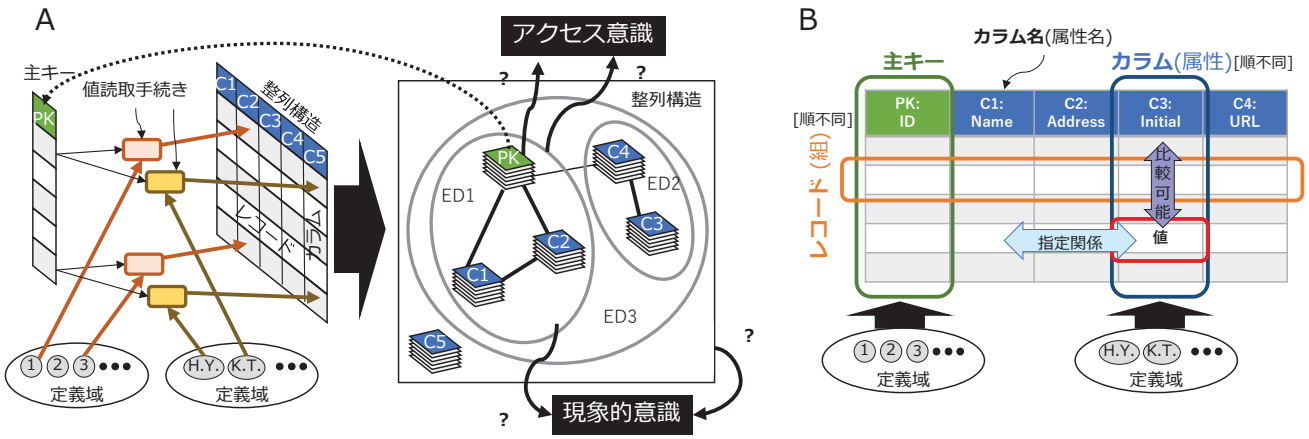

図 1: Entification 概要 (A：Entification 処理、B：整列構造)

B: 括弧内には、RDB の基礎となる関係モデルで使用される名称を記した。

帰納推論を実行できるデータ構造の条件は以下の 3 種類の関係を満たすことである [山川, 2011] [山川, 2021]。それら条件を満たす RDB の構造を整列構造と呼ぶ。

- 指定関係：同一レコード上にある異なるカラム上のフィールドを相互に指定できるという関係。
- 指定関係等価性：レコードを変えても同じ指定関係が成立するという二次の関係。
- 比較可能：同一カラム上にある任意のフィールド間で、少なくとも値の同一性か類似性を判定できる関係。

上記の指定関係とその等価性を自動的に満たすように整列構造を構築する方法が提案された[山川, 2021]。それは図1Aに示すように「**主キーの各値に応じて、特定の定義域から値を読み取る手続きの集合から構築されるデータ構造**」と定義される。なお複数の値読取手続きが定義域を共有可能である。こうした整列構造の定義であれば、例えば、既存の主キーを間引くことで新たな主キーを作成したり、定義域を拡張したり、値読取り手続きに修正を加えたりして新たな整列構造を生成できる。

2.2 Entity 記述

Entity 記述は、「**外界に接地した整列構造上において特定の秩序に支配される範囲を特定できるように記述されたモデル**」である。モデルが整列構造上でコンパクトな記述を持つことで、それを対応付けうる独立した部分情報の範囲を限定できる。

モデルが秩序を記述する形式として、繰り返し現れるパターン、自然変換や同変性を通じて保存されるパターン（準同型）[西郷, 2019] [山川, 2019]、小さな自由度から生成される分布等がある。これらモデルは観測に対する予測の誤差を最小にすることで改善される。

同一の種(kind)とみなされる事物の集合は、それらの内部に含まれる要素間に類似した内部構造をもつ。よって適切な整列構造上における変換 により次の 3 つの情報に分解できる。

- 不変部分：Independent な存在
- 変動部分
  - ➢Dependent な成分：カラム(属性)
  - ➢残差（ノイズ）

図 1A の右側では、互いに密な振る舞いをもつカラムをまとめ上げて Entity 記述を構成する様子を示す。例えば、地球(ED1)を公転する月(ED2)とすれば、そこに含まれる部分を記述するカラム(属性) PK, C1 ~C5 のまま扱うより遥かに少ない記述で全体を概ね把握できる。こうした物理的存在においては、PK,C1, C2 は ED1 の部分であり、ED1、ED2 は ED3 の部分であるといった部分-全体関係がある。こうして、有限な情報（データ量・センサの種類）しかもたない知的システムが Entity 記述を用いることで効果的に複雑な世界を捉えうる。

こうした Entity 記述は、その内外に利点をもたらす。

- Entity 記述内部での利点：
  - ➢ 汎化能力の向上：モデルの予測性の向上
- Entity 記述外部からの利点：
  - ➢ 指示可能：指し示すことができるひとまとまりの情報となる。記号を付与可能。
  - ➢ 記述量削減：定型的な対象として扱うことで内部の詳細な扱いを省略できる。
  - ➢ 思考空間削減：高次概念の操作

我々人類は、上記の利点を組み合わせてより大きな利点としている。人々は多くの高次概念を言語に対応づけられた形で他者から学び、言語を併用した論理的な操作を用いることで、さらに柔軟な思考を効率的に行っていると思われる。

2.3 Entification 処理

Entification 処理は、「**システムが任意の外部に接地した整列構造から得られたデータ上において Entity 記述を認識しようとする処理**」と定義する。生成モデルの観点からは、観測信号との誤差が小さくなる予測信号を生成できる Entity 記述(モデル)を選択(推論)すると同時に、その基盤となる整列構造の探索も行う処理である。

整列構造の探索を並行して行う場合を Dynamic Entification 処理と呼ぶことにする。一旦、よい Entity 記述を含む整列構造が見つかれば、それは再利用されるために固定化されることを Entification 学習と呼ぶ。その学習で得るか所与の事前知識、整列構造に基づいて、あらたな Entity 記述を学習するのが多くの機械学習でありそれを Fixed Entification 処理と呼ぶ。

2.4 意識の Entification 仮説

意識を担いうる系は原理的に、十分な情報保存容量と処理能力を持ち、実際の世界から十分に独立し、互いにほぼ独立した複数の部分から構成されているべきだろう[Tegmark、2017]。

上記の条件を満たすため、ヒトの意識を担う主要な脳器官として、大脳新皮質が含まれることは明らかであろう。情報統合理論や、グローバルワークスペース理論といった主要な意識の理論においても、新皮質もしくはそれと密に連携する視床の関与が前提となっている。そこで、Entification 処理に利用される整列構造や Entity 記述は新皮質上に存在するものと仮定する。

計算機とは異なる、新皮質上における情報表現の重要な性質として、ある時点において表される内容は一つであり、なおかつ同じ表現は複製できないという性質がある。ここではこれを**単表現性**と呼ぶことにする。これは例えば、100 個の「ネコ」を表す局所的な神経集団が存在した場合に、そのうちの、80 個が今見ているネコを表現し、残りの 20 個が昨日観たネコを表現しているといったコーディングは、困難だろうという直観に基づく。

いずれにしても、意識的な帰納推論を行うには、二つ以上の Entity 記述を比較する必要がある。ここで新皮質の単表現性を考慮すると、時刻 tで表現された内容と、次の時刻 $t+1$で表現された内容を比較するというシーケンシャルな処理になる。よって比較を行うために、新皮質上において、Entity 記述を呼び出すことを統括する主キーは集合ではなく系列となるだろう。

ここでは系列として扱われる主キーを、主キー系列と呼ぶこととし、意識を Entification 処理とみなす以下の仮説を提案する。

意識の Entification 仮説：

意識とは、ある主キー系列で制御された複数の読取手続きから得られるデータ構造（整列構造）上において、何らかの秩序に支配される範囲を特定可能なモデルを認識しようとする処理である

こうした Entification 処理としての意識を端的に述べれば、世界において存在と見做せる任意の対象をみつけようとして探索する活動であろう。主キー列の制御下で逐次認識される Entity 記述が意識内容となる。それは、Dynamic Entification と Fixed Entification のいずれにおいても同様であろう。ただし世界には自身の内部も含まれる。

ただし、特定の整列構造上で行われるEntity記述の認識は、少なくとも2つのレベルで自動化されうる。一つは、繰り返されるEntity 記述の認識シーケンスをチャンク化し、主キー系列が時間的に大きなEntity記述を扱うようにした場合である。さらに、初期視覚情報のような処理では、Entity 記述の認識はアーキテクチャ上で並列に実装されるため、単一の主キー系列によって制御される一般的な意味での意識とはみなされない。

3. Entification 仮説の位置づけ

3.1 意識研究地図上での Entification 仮説

最近、[Niikawa, 2020]によって、定義論、現象論、認識論、存在論、価値論についての問いからなる意識研究の地図が提案された。この地図上にEntification仮説を位置づけられれば、多くの先行研究と比較しやすくなるだろう。

(1) 定義論 (Definition):

これは研究の出発点において足並みをそろえるために「意識という語をどう定義すべきなのか？」というものである。典型的な意識経験から共通点を抜き出す実例アプローチと、意識の本質を選び出すことで定義しようとするアプローチがある。

Entification 仮説は、すでに述べたように情報処理として定義されているため、本質アプローチに属するであろう。

既存の定義論からの意識研究としては、自然種であるとか[Bayne, 2021]、実在を自然変換から把握する[西郷, 2019]などがある。これは存在をとらえる働きを意識と見做す Entification 仮説に近い。こうしてみると意識を定義からはじめようとすると、存在にかかわることは避けられないように思われる。

(2) 現象論 (Phenomenology):

ここでの問いは「意識主体からみたとき、意識はどのようなあり方をしているのか？」というものである。

一つの論点は意識の内容と種類だが、Entification 仮説では、外界の事物のみならず思考のメタ認知や自己などの Entity 記述を含むため、内容や種類という点では非常に広い。

次の論点である意識の構造の点で、Entification 仮説は、Entity 記述を見つけ出そうとする処理であるため、多くの場合に何かに注意を向けている志向性があるだろう。ここで着目すべきEntity 記述がまだ世界の中で特定されていない処理段階でも、探索する処理自体は存在するためそこには意識があるとみなす。

(3) 認識論 (Epistemology):【方法論】

ここでの問いは「私達はどのように自己(もしく他者)の意識について知るのか？」というものであり、視点としては自分自身の意識のあり方の分析(一人称)と、他者の意識のあり方についての分析(三人称)があるが、ここでは前者に着目する。

一人称分析の一つは内観である。ここでは、覚醒しているのに存在を探索する活動(つまり Entification)を行っていない状態について研究することが有効であると思われる。一つの可能性は、「無の境地」(雑念を排除し、完全にものに執着しなくなった心持ち)に達した人にインタビューなどを行う方法である。ただし、内観を行う際に自己を認識すると、無でなくなる点に問題がある。

一人称分析を、外部の視点で行うことは、主に神経活動の測定にかかわるので、次項目の存在論で議論する。

(4) 存在論 (Ontology):

ここでの問いは「意識は世界のうちにどのように位置づけられているのか？」というものである。より具体的には、意識のあり方と相関している神経活動を特定し(相関問題)、さらになぜその相関が成立するのかを問うものである(説明問題)。

先に、説明問題に対しては、同一性、随伴現象説、中立一元論、観念論などがあるとされている。現段階で Entification 仮説は、意識を Entification 処理とみなすという、同一説の立場をとっている。その評価方法としては、後述するように意識にかかわる既存の研究やアイディア(クオリア、アクセス意識、自己認識)と整合性を議論することがあげられるだろう。

脳内における意識経験(意識内容)についての主な問いは、次のようなものである。

- 特定の意識内容に十分な最小限の神経活動を特定する
- それらは脳内のどこにあるかを特定する
- そうした意識内容に対して情報的・機能的意味を与える

これらの問いに対する答えは 4 章で述べるような Entification の処理を行えるニューラルネットワークなどの計算モデルの研究開発が進むことで解決に向かうだろう。それは、Entification としての機能を備えたモデルの挙動が、脳の解剖学的構造とタスク実行時の神経活動と対応づけられる研究によるだろう。

(5) 価値論 (Axiology):

ここでの問いは「意識にどのような価値があるのか？」というものである。そこには、システム自身の能力としての機能的価値、認識的価値と、外部からそのシステムをとらえた場合の、道徳的価値、美的価値が想定されている。

Entification仮説においては、システムがEntity記述を獲得することで世界について効率的に記述できるという認識論的価値は明確である。また Entity 記述は、言語的コミュニケーションの基盤となり、システムの行動決定にも役立つため、機能的価値も明確である。

他方で、世界の中の存在をとらえようとする処理に対してどの程度の尊さや美しさといった価値を与えるかは、定義それ自体から導くことは難しいであろう。しかしながら、存在論議論の最後に述べたように、より高い Entification としての機能を発揮できる価値をもつにもかかわらず、異なる実装と振る舞いを示す意識に対して道徳的・美的価値をどのように認めるかという議論の土台を提供できる可能性はある。

3.2 現象的意識とアクセス意識

意識は、必ずしも重複しない現象的意識とアクセス意識に分類されることもある [Block, 2005]。端的には、読書中において、注視点の付近で単語として報告できる部分がアクセス意識で、注視点周辺で報告できないが意識にはのぼっている部分が現象的意識とされる。以下では Entification 仮説においてその両者の対応箇所を検討する(図1参照)。

(1) 現象的意識とそのクオリア

現象的意識(phenomenal consciousness)[1]は質的な内容を伴う主観的な経験と感覚であり内観されるものである。

関連して、典型的なクオリアは、「イチゴのあの赤い感じ」のように物理的対象についての質にかかわる。よって Entification 仮説において、クオリアは Entity 記述のカラム(属性)と関連づくだろう。ここで、さらにクオリアは現象的意識に含まれる個々の対象についての質的な性質であると考えられるようである[2]。そうで

[1] Wikipedia より 「現象的意識」, access 2022/02/28

[2] 脳科学辞典より 「クオリア」, access 2022/02/28

あれば、現象論的意識は、Entification におけるある時点において認識している Entity 記述もしくは、それを含む整列構造に対応づくかもしれない。

(2) アクセス意識

アクセス意識（access consciousness）は、意識の中で明確に注意を向けられている部分である。その意識内容が思考や報告に利用されうるような状態である[太田, 2007]。また各タイムステップで、行動／計画／想像／想起の条件となれる [Bengio, 2019]。

言語報告しうる明示的な意識内容という性質は、Entification 仮説においては主キー系列と関連づけるのが自然である。

しかし、外界からの強い刺激（例えば爆発音）を受けたり、新たなアイディアを探して視点を切り替えようとしたりする際(Dynamic Entification 処理)には、明確に目覚めているが、特定の主キーに拘束されていない。そうした場合を含めるとアクセス意識を単純に主キー系列に対応付ければよいかは疑問が残る。

こうしてみると、Entification 処理の特定の側面を、現象意識とアクセス意識に対応付ける自明な答えはないかもしれない。むしろ今後は、Entification 仮説における諸側面を起点として、それを意識の様々な現象に対応付けることを試みても良いだろう。

4. 脳に学ぶ Entification 機構の探求に向けて

現状の Entification 処理[山川, 2022]では、基本的な機構の提案（図1参照）にとどまっている。今後においてその機構を Neural network 等でモデル化しソフトウエアとして実装を進めることで研究が進展するだろう。そうした過程を経て、3.1 節で述べた存在論的な問いへも次第に答えてゆけるだろう。

Entification 仮説の下では、進展著しい意識とその脳に関わる知見を参照した SCID 法[Yamakawa, 2021]により Entification 機構の機能仮説を設計できるだろう。SCID 法では、脳の主に解剖学的構造を参照しながら体系的な機能の仮説を設計する。

以下では SCID 法の適用にむけた重要なステップである、対象脳領域（ROI）の特定と、それが担う最上位機能(TLF)の特定の検討状況について述べる。まず、ROI として、意識にかかわる脳器官としては、伝統的に新皮質と視床が着目されてきた。しかし、近年は、視床-皮質結合のみに注目した意識障害モデルでは、実験データに適合しないことが指摘され、意識における基底核の関与が有力視されている[Crone, 2017]。次に TLF は、意識の Entification 仮説を反映し、「様々な整列構造上における、何らかの秩序に支配される範囲を特定可能なモデルを獲得し認識する機能」として定義するのが妥当であろう。

上記の準備のもとに SCID 法を進めてゆくにあたり、意思決定にかかわる基底核は、主キー系列を制御する機能を担う機構として自然に想定可能なため、脳に適合した機能仮説の構築にむけて貴重な足掛かりとなりうる。他方で Entification 処理の構成要素である定義域をモデル内で実現するために、自動的に相互に比較可能な値を取得できる仕組みの検討が待たれる。

5. おわりに

本稿では、外界からの情報を様々な Entity 記述に変換する Entification 処理において、主キーが系列化したものが意識であるとする Entification 仮説を作業仮説として設定した。次に作業仮説を意識研究地図上で位置づけ現象的意識とアクセス意識との関連を議論した。さらに Brain-inspired な Entification 研究の可能性について検討した。

今後、脳を参照しながら意識としての Entification のモデル化を試みる。その際にはメタ認知や自己認知といった認知機能や、海馬などの脳器官との関連[山川, 2020]も考慮すべきであろう。

ところで、Entification 仮説を前提とするならば、AI はいずれ、人とは異なる形で人を超えた意識のようなものを有するだろう。その時に AI も存在を見出そうと努力するが、そのプロセスは単一の系列ではない点が人とは異なる。そうなると「AI に意識が宿るか？」という問いはより複雑な様相を呈するかもしれない。

最後に一言加えるなら、実は、我々にとってあらゆる煩悩を消し去った、「無の境地」に到達することが難しい理由は、実は、意識というものが存在を探し求める Entification であったこと反映しているのかもしれない。

謝辞： 本研究にあたり、松尾研究室内にて折に触れて議論頂いている鈴木雅大氏、岩澤有祐氏、熊谷亘氏、に深く感謝する。

参考文献

[山川, 2022] 山川宏, Entification の理論を目指して: 世界から存在を取り出す一般的な原理とは, 人工知能学会, SIg-FPAI, 2022.

[Codd, 2002] Codd, E. F. A Relational Model of Data for Large Shared Data Banks. in Software Pioneers: Contributions to Software Engineering (eds. Broy, M. & Denert, E.) 263–294 (Springer Berlin Heidelberg, 2002).

[山川, 2011] 山川宏. 脳から学ぶべき知的能力は何か. 2011 年度人工知能学会全国大会（第 25 回）, JSAI2011 2C2-OS2b-4 (2011).

[山川, 2021] 山川宏. 客体化学習の検討.人工知能学会全国大会（第 35 回）1H2-GS-1a-01 (2021).

[西郷, 2019] 西郷甲矢人 & 田口茂. ‘現実’とは何か: 数学・哲学から始まる世界像の転換. (筑摩書房, 2019).

[山川, 2019] 山川宏. 圏論からみる実在の仮説. 日本認知科学会第 36 回大会 OS01-4 (2019)

[Tegmark, 2017] Tegmark, M. Life 3.0: Being Human in the Age of Artificial Intelligence. (Knopf Doubleday Publishing Group, 2017).

[Niikawa, 2020] Niikawa, T. A Map of Consciousness Studies: Questions and Approaches. Front. Psychol. 11, 530152 (2020)

[Bayne, 2021] Bayne, T. The Unity of Consciousness. (Oxford University Press, 2012).

[Block, 2005] Block, N. Two neural correlates of consciousness. Trends Cogn. Sci. 9, 46–52 (2005)

[Crone, 2017] Crone, J. S., Lutkenhoff, E. S., Bio, B. J., Laureys, S. & Monti, M. M. Testing Proposed Neuronal Models of Effective Connectivity Within the Cortico-basal Ganglia-thalamo-cortical Loop During Loss of Consciousness. Cereb. Cortex 27, 2727–2738 (2017)

[太田, 2007] 太田,紘史, 意識の表象理論. 哲学論叢 2007, 34: 102-113, 2007.

[Bengio, 2019] Yoshua Bengio, Challenges towards AGI, Beneficial AGI conference ,2019.

[Yamakawa, 2021] Yamakawa, H. The whole brain architecture approach: Accelerating the development of artificial general intelligence by referring to the brain. Neural Netw. (2021)

[山川, 2020] 山川宏. 内観における Papez 回路の役割について. 人工知能学会第二種研究会資料 2020, 03 (2020)