深層ニューラル・ネットワークと概念所有

エピソード記憶を用いた深層強化学習モデルに焦点を合わせて

西村正秀
Seishu Nishimura
滋賀大学 経済学部 / 教授

I はじめに

現在、深層学習をベースとしたAIは私たちの生活に欠かせないものとなっている。AIが認知モデルとして用いるのはコネクショニズム（ニューラル・ネットワーク）である。1950年辺りに登場したコネクショニズムは、1980〜1995年に大きなブームを迎えた。その後一時期停滞したが、2010年代以降再びブームとなっている。現在のコネクショニズムと旧来のコネクショニズムの違いの一つは、前者において隠れ層が多層化された（「深くなった」）点にある。このような技術的進歩がコネクショニズムに新たな性質を付与するのか否かについては、現在哲学的評価が始まった段階だと言えるであろう。

本稿では、現在のコネクショニズムにおいて概念所有は十分な仕方で説明されているのかという問いを考察する。概念とは何か、また、概念所有の条件は何かという問いは、哲学のみならず、心理学や認知科学といった経験科学において長く議論されている。哲学とこれら経験科学でまったく同じ問いが扱われてきたのかは疑わしいが（Machery (2009)）、両分野で概念が主題となる際に、「思考の体系性（systematicity）」がどのように説明されうるのかという論点が中心とされてきたことには間違いはない（Camp (2009)）。思考の体系性とは、「任意の表象（文）からその背後にある統語論的構造に応じて意味を再構成できるという特性」のことである（谷中・峯島 (2024), 28）。議論のために、思考は心的表象だと仮定しよう。例えば、「太郎は花子を愛している」という思考を「太郎」「花子」「〜は…を愛している」という要素に分解し、そこから「花子は太郎を愛している」という新しい思考を再構成できる主体は体系性を持つ。旧来のコネクショニズムは、この体系性を十分に説明する

ことができないと診断されてきた。例えば、M. Davisによれば、コネクショニズムが措定する分散表象では特定の概念（例えば「コーヒー」）について同定できる統語論的ユニットを形成できず、精々家族的類似性のパターンしか見出すことができないので体系性は説明できない（Davis（1991））。これらの診断者によれば、思考の体系性は表象の統語論的構造を確保する「思考の言語（以下、LOTと略記）」仮説を中核とした古典的計算主義だけが説明できる性質なのである。

議論のために、旧来のコネクショニズムに対する否定的診断は正しいとしよう。現在のコネクショニズムでは「深層ニューラル・ネットワーク（以下、DNNと略記）」が用いられる。さらに、DNNの学習についても新しい方法が提案・実践されている。このような違いに訴えながら、R. MillièreはDNNでも思考の体系性を説明できるという議論を展開している（Millière（2024））。一方、J. Quilty-Dunnらのように、依然としてLOT仮説を擁護する議論も存在する（Quilty-Dunn et al.（2023））。本稿では、Millièreの議論に焦点を合わせて、DNNが思考の体系性を十分に説明できるのか、さらに、概念所有の条件を十分に説明できるのかという点を批判的に検討する。議論は以下の順で行う。最初に、DNNと旧来のニューラル・ネットワークの違いを簡潔にまとめておく（第2節）。次に、思考の体系性と概念所有について説明する（第3節）。続いて、コネクショニズムのライバル理論と見なされてきた古典的計算主義とLOTについて簡単に見ておく（第4節）。その後に、Millièreが現在のDNNで体系性がいかに説明されうるのかを論じる際に言及するN. Sheaの議論を概説して（第5節）、本当にSheaの議論が思考の体系性と概念所有の条件を統語論的構造に訴えることなしに説明できているのかを検討する（第6節）。

II DNNと旧来のニューラル・ネットワークの違い

DNNの特徴を説明する前に、旧来のコネクショニズムとは何かを黒柳et al.（2003）とMillière（2024）を参照しながらまとめておく。ニューラル・ネットワークとは、脳の神経回路網を模した認知モデルである。脳ではニューロンで活性化した電位が情報の担い手であり、ニューロン間の電位の伝達によって認知活動が説明される。電位の伝達量はシナプスの状態で決まる。詳細は省くが、ニューラル・ネットワークでは、シナプスにおける結合の度合い（重み）とニューロンに対する入力から、出力を計算する数学的モデルが立てられる。そして、大量のニューロンを結合させたネットワークによって、多様な認知活動の説明が行われる。このネットワークは、入力層と出力層と両者の間に存在する隠れ層からなる。各層は「ノード」と呼ばれる処理ユニットから構成され、ノードを介して各層は入力を出力にマッピングする。1980年代に入るとニューラル・ネットワークに、分散表象、全結合型ネットワーク、バックプロパゲーション（誤差逆伝播法）といった新しい技術が追加された（Millière（2024），2）。分散表象とは、ネットワーク全体に分散した表象の在り方である。コネクショニズムでは対象を表象する際に、それに対応する特定のユニットは配置されない。その代わりに、ネットワーク上のユニットに対する重みを変化させることによって様々な対象が表象される。全結合型ネットワークとは、各ニューロンが別のニューロンと非線形的に結合したネットワークであり、よりダイナミックな出来事の表象を可能とする。バックプロパゲーションはネットワークに関する教師付き学習方法の一種であり、ある入力に関して目標とされる出力値と実際の出力値の誤差を、

各層に逆向きに伝播させることによって重みを調整するアルゴリズムである。

これらの特徴に加えて、DNNではいくつかの新しい特徴が追加される。C. Bucknerによれば、DNNは旧来のニューラル・ネットワークと次の四点で異なっている（Buckner (2019), Sec. 2）。第一に、旧来のネットワークは隠れ層が3〜4層から構成されている点で「浅い」。一方、DNNは5から250層の隠れ層を有し、その意味で「深い」。この深さによって、DNNは浅いネットワークと比べて飛躍的に高い効率性を有する。第二に、旧来のネットワークは一種類のノードしか含まないが、DNNは畳み込みノードやプーリング・ノードなど異なる活性化機能を持つ多種類のノードが含まれる。第三に、旧来のネットワークでは、ある層のノードはすべてその前の層のノードと結合していたが、DNNでは多くの場合、ある層のノードはその前の層のノードの中で時空的に近いものとだけ結合しており、そのおかげで学習しなければならないパラメーターの数が押さえられている。第四に、DNNでは過学習を避けるために、明確な正則化技術が使用されている。

これらの新しい特徴によって、DNNの説明力は飛躍的に高まった。ただし、DNNは進化し続けており、以下で見るように、思考の体系性や概念所有の説明については、さらなる特徴に注目する必要がある。

III 思考の体系性と概念所有

3.1. 思考の体系性

思考の体系性とは、上述したように、私たちの認知活動に見られる、ある思考の構成要素を結合して別の思考を再構成できるという特性のことである。「太郎は花子を愛している」という思考Aから「花子は太郎を愛している」という思考Bを再構成できる主体は、この再構成を無秩序に行っているのではなく、あるパターンに従って行っている。思考Aの内容は「太郎」「花子」「〜は…を愛している」という要素から構成されるが、各要素には特有の統語論的機能があり、その機能に従ってそれらを適切に再合成することによって思考Bが形成されている。このような各構成要素の機能を把握しているからこそ、私たちは同じカテゴリーの別の要素を用いて「トムはキャサリンを愛している」といった思考を作ることもできる。E. Campが指摘するように、私たちの認知活動におけるこのような体系的パターンは、思考内容が「構造化されている（structured）」という仕方でも表現される（Camp (2007), 147; e.g., Crane (1992)）。

また、思考の体系性と関連する思考の特性として「推論の体系性」も存在する（谷中・峯島(2024), Camp (2007)）。思考内容に体系的パターンがあることは、その思考を用いて主体が多様な推論を行うことができるという事実を説明してくれる。例えば、A∧Bという思考を持つ主体はその思考から思考Aや思考Bを演繹的に推論することができる。この推論は、A∧Bという連言的な思考内容とAやBという思考内容の関係に裏付けられている。このようなパターンに応じて様々な推論が行えるという特性が推論の体系性である。

これらの体系性は人間の認知活動に観察される事実である。したがって、DNNが人間の認知モデルとして説得力を有するためには、これらを十全に説明できるものでなくてはならない。

3.2. 概念と概念所有

「概念とは何か」という問いは大きな問いであり、多様な解答が存在する。本節では、「概念の存在論的身分は何か」という問いと、「概念を所有する

ための条件とは何か」という問いに分けて、本稿で必要な説明を与えておく。

まず、概念の存在論的身分については、本稿では「概念＝心的表象」という仮定のもとで話を進める。一般的に概念としてイメージされるのは「トラ」や「コーヒー」のような種を表す語彙的な表象であろう。「トラ」が「トラは哺乳類である」という思考に登場できるように、心的表象としての概念は思考の構成要素である。このような概念がどのような形而上学的本性を持つのかについては、いくつかの考え方がある。心理学において多くの場合、心的表象としての概念は長期記憶に蓄えられた「知識」と見なされる（Machery（2009））。ここでの「知識」とは認識論で使用される「正当化された真なる信念」ではなく、推論やカテゴライゼーションなど高次の認知活動に使用される構造化された情報の集合を指している。例えば、「トラ」の概念は性質（黄色と黒色の縞がある、など）、行動（肉食である、など）、起源（〜から進化したネコ科の動物である、など）を内容とした表象であり、虎に関する推論やカテゴライゼーションに使用される。また、別の考え方としては、心的表象としての概念を、構造を持たない統語論的に原子的な表象と見なす方法もある（e.g., Fodor（1998））。これらの心的表象としての概念は古典的計算主義でもコネクショニズムでも措定可能なものだが、それがどのような仕方で実装されるのかについては各理論によって異なる。DNNにおいて概念がどのような心的表象として特徴づけられるべきかについては、本稿第6節でより詳細に議論する。

続いて、「概念を所有するための条件とは何か」という問題に移ろう。この問題については、条件に何を数え上げるかに応じて異なる「概念」概念が提案される。多くの論者がミニマムな条件として数え上げるのは、G. Evansが提唱した「一般性制約（the Generality Constraint）」である（Evans（1982），104）。

> もしある主体に「aはFである」という思考を帰すことができるならば、その主体が概念を有している「Gである」という性質すべてについて、その主体は「aはGである」という思考を享受する概念的リソースを持っているに違いない。（*Ibid.*）

これは、ある主体がある概念を所有していることを、その主体がその概念を他の性質や個体にたいして一般的に適用する能力を有していることに、すなわち、表象を体系的に分解・再合成する能力を有していることに紐づけようという主張である。「表象内容は構成要素に分解して再合成できる」という思考の体系性は、表象内容を操作する能力を必要としている。一般性制約は、その能力を概念所有の必要条件と見なすアイデアである。本稿でも、一般性制約を概念所有の必要条件と見なすアイデアを採用する。

では、一般性制約に加えて何が概念所有の条件として数え上げられるのか。この問題については論者によって解答が分かれる。Campはこの解答を「ミニマリズム」「穏健な立場」「知性主義」の三つに分類している（Camp（2009））。ミニマリズムとは、一般性制約を概念所有の必要十分条件とする立場である。穏健な立場は、一般的制約に加えて「状況からの独立性」を概念所有の必要条件と見なす立場である。「知性主義」は、穏健な立場の条件に、言語を使用できることという条件を追加する立場である。知性主義はR. Descartes以来の立場であり、現代でもD. DavidsonやJ. McDowellなどが擁護しているが、人間以外の生物が概念を所有する可能性を閉ざす帰結を伴って

おり、動物の認知に関する様々な経験科学の知見と相反する。また、ミニマリズムは、概念的思考において主体は、たとえ表象される対象が存在していなくても、その対象について思考することができる（例えば、目の前に犬がいなくても、犬について考えることができる）といった「表象と表象されるものの明確な区別」あるいは「状況からの独立性」を十全な形で説明することができない。以上の理由から、Campは穏健な立場を最も有力な解答として評価している。本稿でも、作業仮説として穏健な立場を使用する。

なお、ここで注意が必要なのは、概念所有は程度問題だという点である（Camp（2009））。一般性制約が求める表象の一般的適用能力には、もちろん個人差がある。たとえ局所的な仕方でしか適用できない場合でも、それに基づいて有用な仕方で自分の行動を導くことができれば、主体はその概念を低い程度で有していると言える。この点は、以下でDNNが概念所有を十全に説明できるか否かを検討する際にもう一度触れることになる。

Ⅳ 古典的計算主義と思考の体系性

本稿の目的は、DNNが思考の体系性や概念所有を十分に説明できるか否かを検討することであるが、その際に対比軸となるのは古典的計算主義である。というのも、1980〜1995年のコネクショニズムがこの問題に関して否定的診断を受けてきた背景には古典的計算主義との対比があり、DNNの評価もそれが古典的計算主義に対する不足点を乗り越えることができているのかという観点からなされるべきだろうからである。具体的に言えば、旧来のコネクショニズムに対しては、(1) 思考の体系性を十分に説明できないか、(2) 仮にそれを説明できるとしても、それは結局古典的計算主義が前提するLOTを実装することによってであるという評価が下されてきた（e.g., Fodor and Pylyshyn（1988）；Millière and Buckner（2024））。これを「体系性のジレンマ」と呼ぼう。本稿におけるDNNの評価も、DNNがこのジレンマに陥っていないかどうかという観点からなされる。本節では、そのために古典的計算主義ではなぜ思考の体系性をうまく説明できるのかを簡単に見ておきたい。

まず、古典的計算主義という考えを簡単に説明しておこう。これは私たちの認知を心的表象の統語論的操作によって説明する理論である。戸田山和久によれば、古典的計算主義は、(1) 機能的分解、(2) 統語論的構造を持つ表象、(3) アルゴリズム的な情報処理という三つの特徴を持つ（黒柳et al.（2003），14-15）。(1) は、(i) 心は顔認識、文字認識などの機能を担ういくつかのサブシステム（モデュール）に分解される、(ii) それらのサブシステムはより下位のサブシステムに分解される、(iii) それら下位のサブシステムもより下位のサブシステム（例えば、知覚レベルのサブシステム）に分解されるといったように、心を階層的な仕方で説明するアイデアである。(2) は、それら個々のサブシステムにおける情報処理は、LOTと呼ばれる統語論的構造を持った一群の心的表象を用いて説明されるという主張である。(3) は、LOTを用いた情報処理は予め与えられたアルゴリズムによってなされるという主張である。このアルゴリズムによる心的表象の統語論的操作が「計算」と呼ばれているわけである。

古典的計算主義者によれば、「古典的計算主義が街で唯一のゲーム（the only game in town）である」（Quilty-Dunn et al.（2023））というスローガンに表されるように、LOTとして同定された表象体系の統語論的操作に訴えることによってのみ、

思考の体系性は説明できる。Campは、LOTだけが思考の体系性を説明できることを結論付ける論証を次のようにまとめている（Camp（2007）, 146;「思考の言語」はLOTという表記に筆者が統一した）。

1. 思考主体が表象し推論できる諸内容間には体系的関係がある。
2. 諸内容間の体系的関係は、思考主体の表象的かつ推論的能力における相関的構造に反映されていなければならない。
3. 構造化された表象能力は、体系的規則によって合成される再帰的で離散的な部分からなる表象媒体の体系を要求する。
4. 体系的規則によって合成される再帰的で離散的な部分からなる表象媒体の体系は言語に他ならない。

それゆえ、LOTは存在しなければならない。

この論証を「唯一のゲーム論証」と呼ぼう。この論証のうち、前提1と2は前節で確認した通りである。前提1は思考や信念を表象的な心的状態として理解する以上、私たちの認知活動に関する事実から導き出される主張である。前提2は、思考の体系性を説明するためには、単なる表象メカニズムだけではなく、様々な思考や推論を可能とするように、表象内容の諸部分をそれぞれ一般的な仕方で運用する能力が要請されるという主張である。この主張は、各思考をバラバラに説明するよりも、そのような能力を措定することが前提1で見られる現象を最も効率よく説明できるという理由で擁護される（Fodor and Pylyshyn（1988））。前提3は、前提2で措定される能力は、表象媒体の統語論的操作によって遂行されるという主張である。古典的計算主義者によれば、表象内容を構成要素に分解し、それを再構成するためには、その内容が因果的効力を有した物理的な媒体によって実装される必要がある。別の言い方をすれば、表象内容と媒体の同型性が必要だというわけである。前提4は、統語論的構造を持つ表象媒体は言語（LOT）という形しか取りえないという主張である。

この論証は妥当である。したがって、この論証を反駁するためには、前提の少なくとも一つを否定しなければならない。どの前提を否定するかについては、いくつかの可能性がある。例えば、分散表象を措定するコネクショニズムは、前提3を否定する理論だと解釈できる。ただし、前提3をどのような仕方で否定するのかについては、さらにいくつかの可能性がある（消去主義をとる論者もいれば、より穏健な折衷案をとる論者もいる）。また、後で見るように、前提4を否定する立場（認知地図理論など）も存在する。それゆえ、古典的計算主義が「街で唯一のゲーム」であることは疑わしい。実際、現在LOTを擁護しているQuilty-Dunn et al.（2023）も、それが「街で唯一のゲーム」だという主張から「街で最良のゲーム（the best game in town）」だという主張に後退している。とはいえ、仮にLOTの存在を認めれば、それが前提1〜4を説明する有力な仮説の一つとなることは明らかであろう。以下では、この論証をどのようにDNNが崩していけるのかという問いに対する一つの答えとして、Millièreの議論を吟味していく。

V エピソードDNN

Millièreの議論も「唯一のゲーム論証」の前提3を否定するものである。彼の議論は多岐に渡るが、その中で核となるのは、DNNでは古典的計算主義とは異なる計算メカニズムが実装されているという議論である。思考の体系性にとって必要な特

徴に、「役割と充填物の独立性（the role-filler independence）」がある。これは「離散した記号表象に対して変数束縛を行う能力であり、そこでは「役割」（変数）と「充填物」（値）は独立に表象される」（Millière（2024），4）。例えば、「〜が…を愛している」の「〜」の役割は名詞によって担われるが、そのことと名詞として具体的に「トム」が入るか「太郎」が入るか（あるいはその他の表現が入るか）は独立の事柄である。古典的計算主義がこの特徴を持つことは、LOTの統語論的構造から見て明らかであろう。一方、DNNもこの特徴を有することは経験科学的に示唆されているが、それはベクトル部分空間を用いたメカニズムによって実装されている。このメカニズムは、LOTが用いる記号表象と機能的には正確に対応しない「ファジー」なものであり、その結果、DNNにおける役割と充填物の独立性は程度を許容するものとなる（*Ibid.*）。これは思考の体系性について、古典的計算主義とDNNが異なる計算メカニズムを実装していることを示唆している。

この議論を補強するためにMillièreが持ち出すのが、「DNNにエピソード記憶を付加する」というSheaが取り上げたアイデアである（Millière（2024），4; Shea（2021））[1]。従来のDNNや深層強化学習には、大量の試行の必要性というデータ効率の悪さや、出力が変化すると再学習が必要になるという柔軟性の無さなどの制約があった。Botvinick et al.（2019）を参照しながら、Sheaはこの制約を解消するためにはメタ学習とエピソード記憶の追加という二つの方法があると論じる（Shea（2021），159）。前者は学習の仕方を学習させる方法であり、特定の入力―出力マッピングについて、DNNにその問題と関連する多くの人工的なタスクを与えて、その問題に関する一般的知識を獲得させる。この知識は明示的な合成規則を用いずに獲得される（Millière（2024），3）。この知識によって、DNNは新しい環境に出会ったときに自分を素早く適応させていくことが可能となる。一方、後者はDNNにエピソード記憶、すなわち、「以前に出会った状態と報酬の明示的表象」という個別的エピソードを付加することで制約を解消する方法である（Shea（2021），160）。例えば、ビデオゲームの仕方を学習するDNNでは、各状態は画面上のピクセル配置の表象であり、報酬はゲームの勝利である（*Ibid.*）。Sheaによれば、DNNはある状態に遭遇したときにその都度生じることを、それぞれ別のエントリーとして記録できる。また、ある状態に遭遇したときにとった行為とその行為によって獲得した報酬も記録できる。これらの記憶表象に基づいて、DNNは新しい状態に出会うと、過去の類似した状態でとった全行為の平均値を、それらに現在の状態と似た重みを与えて計算し、現在とりうる行為の値を算出する（*Ibid.*）。この計算を行う際に学習が必要なパラメーターの数は固定されておらず、記憶に蓄えられるエピソードが増えるほど、現在とるべき行為の計算も複雑になる。だがこの方法を用いれば、新しい状態を一つ与えるだけで計算できるという、いわゆる「One-shot学習」で問題を解決できる。また、貯蔵されたエピソード記憶はオンラインだけではなくオフラインでも使用できるので、過去の経験をリプレイして、直接遭遇したことがない異なる刺激同士の相関を学習できるという強みもある。エピソード記憶を用いた深層強化学習モデルを「エピソードDNN」と呼ぼう。メタ学習が思考の体系性を説明できるのかについては、いくつかの先行研究がある（Millière and Buckner（2024）; Millière（2024））。本稿では、エピソードDNNに焦点を合わせて、それが思考の体系性を十全に説明できるのかを見ていく。

1）なお、私が読む限り、Shea自身はエピソードDNNが思考の体系性や概念所有を非統語論的な仕方で説明できるという主張はしていない。

Sheaによれば、旧来のニューラル・ネットワークやDNNと比べてエピソードDNNは二つの利点を持つ。一つは、記憶を計算と明確に分離して明示的表象として蓄えることができる点である（Shea（2021）, 161-164）。これは記憶に蓄えることができるエピソードを自在に増やせるだけでなく、何らかの干渉を受けてもそのまま記憶を保存できることを意味する。また、記憶と計算を区別しなかった旧来のニューラル・ネットワークと比べて、エピソードDNNでは情報処理が遥かに容易なものとなる[2)]。もう一つの利点は、「タスク変数の特定の値から独立したアルゴリズムを使用する」ことができる点である（*Ibid*., 164-167）。より簡単に言えば、これは入力が何であれ、同じ仕方で計算ができるということである。Sheaはエピソードの計算を「内容特定的計算（content-specific computation）」と「非内容特定的計算（non-content-specific computation）」の二つに区別する。前者は、入力と出力が真理保存的関係を持つ（真なる表象が入力されると真なる表象が出力される）ことが、入力と出力それぞれの特定の表象内容に依存している計算である。それに対して、後者はタスク変数の特定の値から独立したアルゴリズムを用いた計算であり、入力と出力の表象内容が何かを問わずになされる。新しい状態に出会った際、エピソードDNNは「蓄えられている記憶をすべて思い出して、その内容に関わらず、それら記憶と新しい状態の類似性を同じアルゴリズムを用いて計算する」（Millière（2024）, 4）[3)]。非特定内容計算は、量化を用いた論理的推論などに対応するものである。

旧来のDNNは特定内容的計算しかできないが、エピソードDNNはそれに加えて非特定内容的計算もできることが強みである。これが強みであるのは、非特定内容的計算は、古典的計算主義の利点であったより抽象的な統語論的操作をカバーしうるからである。古典的計算主義では、上述した「役割と充填物の独立性」が担保されているため、例えば量化などの束縛変項を用いた計算も容易に行える。この利点は特定内容的計算ではカバーできないが、非特定内容的計算ならば同様の抽象的計算が可能となる。さらに、非特定内容的計算は、xやyといった変項に対してだけでなく、xやyに特定の内容を持つ表象が代入されている事例にも適用可能である。例えば、次の二つの事例を見てみよう（Shea（2021）, 165）。

（1）すべてのスズメは飛ぶ。エイブはスズメである。それゆえ、エイブは飛ぶ。

（2）$(x-1)^2 \equiv x^2 - 2x + 1,$
$\therefore 2x \equiv x^2 + 1 - (x-1)^2.$

（2）は変項を含むが、（1）では変項ではなく「スズメ」「エイブ」「〜は飛ぶ」といった特定の内容を持つ表象が使用されている。だがこれらの代わりに「鯉」「トム」「〜は泳ぐ」が使用されていても、同じ計算（推論）パターンが遂行される。これは、非特定内容的計算が古典的計算主義で使用される計算よりも一般性が高い抽象的計算であることを意味している。

このようなエピソードDNNにおける内容特定的計算と非内容特定的計算の区別は、思考の体系性を説明する可能性を開く。Sheaによれば、エピソードDNNは「合成性（compositionality）」を有する（Shea（2021）, 171-172）。合成性とはある思考を分解して新しい思考を再構成するという性質であり、ここでは便宜上、思考の体系性と同義のものとして扱っても構わない。エピソードDNNと合成性の関係は複雑である。内容特定的計算だけが可能なDNNにおいても、「イメージから対象や性質を抽出する」レベルでの合成性は確認されて

2）旧来のニューラル・ネットワークでは、過去の記憶はすべて特定の入力―出力の連鎖によって保存され、現在の状態を推定する際には、可能な過去の状態の連鎖の組み合わせを、重みづけを調整しながらすべて計算しなければならず、極めて負荷の高いメカニズムとなっていた。

3）非内容特定的計算ができることと明示的な記憶表象を持つことの間に必然的な関係はない（Shea（2021）, 164-65）。

いる（Shea（2021）, 171）。一方、記憶に蓄えられる明示的表象は、原理的には分解・再構成可能な構造を持たなくてもよい。むしろ、重要なのは、非特定内容的計算がLOTにおいて可能であった思考の体系性（と推論の体系性）をカバーしうることである。非特定内容的計算は、上の（1）で見たように、文のような合成的表象を入力とする一般性の高い推論を可能にする。（1）において「スズメ」を「カラス」と変えても同じ推論パターンが遂行されることは、構成要素を入れ替えて新しい思考や推論を形成することができるという体系性を担保するものである。ここで重要なのは、非特定内容的計算は思考の体系性（と推論の体系性）をLOTとは異なるメカニズムで補っていると考えられることである。Millière によれば、エピソードDNNでは特定内容的計算と非特定内容的計算は厳密に区別されておらず、計算の内容特定性は「程度問題」である（Millière（2024）, 4）。上述したように、この「ファジーさ」は古典的計算主義にはなかった性質であり、その点で非特定内容的計算はエピソードDNNがLOTとは異なるメカニズムを有することを傍証している。

以上が、エピソードDNNは思考の体系性をLOTとは異なるメカニズムで十分に説明できるという議論である。では、エピソードDNNは概念所有については何が言えるのか。Sheaも「概念＝心的表象」という構図の元でエピソードDNNを論じている。SheaはCampとEvansに言及しながら、概念は推論や熟慮した思考に登場できるだけではなく、「新しい思考を生成するために相互に自由に再結合できる命題の下位構成要素（sub-propositional elements）」であると特徴づけている（Shea（2021）, 167）。彼によれば、このような概念は内容特定的計算だけではなく非内容特定的計算にも登場することができる（*Ibid*., 168）。後者の一例である上の事例（1）では「スズメ」などの概念が登場しているが、推論パターン自体は特定の概念内容とは関係なく機能している。概念が非内容特定的計算に登場するという事態は、非内容特定的計算が思考や推論の体系性をカバーするものであったことを勘案すれば、「穏健な立場」が概念所有に要求する条件、すなわち、一般性制約と状況からの独立性を実現していると解釈することができる。

だが、このような概念所有の条件を実際に説明するためには、どのようなモデルが必要とされるのであろうか。この解答として、Sheaは二つの候補を挙げている（*Ibid*., 168-169）。一つはHummel and Holyoak（2003, 2005）による「「可能性（how possibly）計算モデル」であり、これは計算主義とコネクショニズムのハイブリッドモデルである。もう一つは、Eliasmith（2013）やQuilty-Dunn（2021）が提案するポインター・モデルである。細部に違いはあるが、彼らの議論は概念を「ポインター」あるいは「指示詞的表象」と呼ばれる心的表象と見なす点で一致している。ポインターは特定の内容が保存された記憶の場所を指示する抽象的表象であり、原子的表象の一種である。このモデルにおいては、特定の内容、すなわち、「知識」（推論やカテゴライゼーションに使用される情報の集合）は、「concepts（概念）」ではなく「conceptions」と呼ばれる。ここでポイントとなるのは、ポインター自身に計算が適用されるときには、それが指示する場所に記憶として保存されている特定内容が何であるかは考慮されないということである。もし概念をポインターと見なすことが可能ならば、この道具立てで非特定内容的計算を説明できることは明らかであろう[4)]。

以上の議論をまとめておこう。Millièreによれば、思考の体系性はDNNによって十分に説明できる。

4）厳密に言えば、概念は単なるポインターではなく、生成ポインター（generative pointers）である（Quilty-Dunn（2021））。概念には「chicken」のように多義的なもの（鶏の意味も鶏肉の意味もある）がある。これらの特定内容表象（conceptions）は同じ記憶に蓄えられており、どちらの対象を表示するかはポインターだけでは決定できない。そこで、ポインターは一部の特定内容表象と結びつくことによって特定の対象を表示する。このように指示を生成するという意味で、概

非統語論的メカニズムを具体的に表すモデルの一つがエピソードDNNであり、そこで肝となるのは非特定内容的計算である。この計算は表象として蓄えられたエピソード記憶と新しく与えられた状態との類似性の比較に基づくものだが、記憶表象の特定の内容は無視して一般的な推論パターンのみに着目する。このような一般的推論によって、古典的計算主義においてLOTがなしていた統語論的分解・再構成に基づく体系性はDNNにおいてもカバーすることが可能となる。さらに、非特定内容的計算は、主体が概念的思考を行う際の能力を担保するものとしても機能している。

VI エピソードDNNは思考の体系性と概念所有を十分に説明しているのか

だが、前節で見たエピソードDNNは、本当に思考の体系性ならびに概念所有を十分な仕方で説明しているのであろうか。本節ではこの問題を検討する。結論を先に述べれば、エピソードDNNはコネクショニズムに従来指摘されてきた「体系性のジレンマ」を克服できず、思考の体系性と概念所有を十分な仕方で説明できないと考えられる。以下、その理由を説明しよう。

エピソードDNNが思考の体系性を説明する際に決定的な役割を果たしていたのは、非内容特定的計算であった。Sheaは非内容特定的計算が行われる典型例として、概念的思考を取り上げた。問題は、この計算の対象となる「概念」がどのように説明されているのかにある。この説明モデルとして、Sheaは「可能性計算モデル」と「ポインター・モデル」の二つを紹介していた。この内、計算主義とコネクショニズムのハイブリッドである前者は結局のところ思考の体系性の説明をLOTに担わせるものであり、本稿第4節で触れた「体系性のジレンマ」の第二の角に当てはまってしまう。それゆえ、より見込みがあるのは後者のポインター・モデルとなる。

では、ポインター・モデルは概念所有に関する説得力のある理論なのであろうか。ここで問題として考えられるのは、ポインター・モデルはコネクショニズムではなく、古典的計算主義の方と相性がよいという点である。実際、ポインター・モデルは原則的にLOTと紐づけられている（e.g., Quilty-Dunn（2021）; Murez（2021）; Coelho Mollo and Vernazzani（ms））。その点を、Shea自身が言及していたQuilty-Dunnのポインター・モデルを例にとって解説しよう（Quilty-Dunn（2021））。Quilty-DunnはM. J. Greenと共にポインター・モデルを展開している（Green and Quilty-Dunn（2021））。彼らによれば、概念はポインターと呼ばれる原子的な心的表象である。このアイデアの背後にあるのは、「対象ファイル（object files）」（あるいは「心的ファイル（mental files）」）という道具立てである[5)]。元来、対象ファイルは心理学において知覚システムの説明に使用されてきた道具立てであったが、現在は単称指示や単称思考（singular thoughts）を説明するために哲学でも用いられている（e.g., Recanati（2012）; Green（2018））。対象ファイルの本性はまだ完全に明確にはされていないが、多くの論者が共通見解として挙げているのは次の二つの特性である。

> 対象ファイルは一般的に、(i) 外的対象への通時的な指示を確保し、(ii) その対象の性質に関する情報を保存してアップデートする表象として特徴づけられる。（Green and Quilty-Dunn（2021）, 666）

念は生成ポインターと呼ばれるわけである。

5）心的ファイルは、対象ファイルをより一般化して高次認知システムにも拡張した道具立てだと言える。心的ファイルの説明については、Murez and Recanati（2016）を参照されたい。

このように、対象ファイルは知覚における対象への指示と、その対象を記憶(ワーキング・メモリー)に保存されている情報と結びつけるという二重の役割を果たすものとして想定されている。これら二つの特性が対象ファイルに帰せられる理由は、心理学における対象再認実験の結果(被験者に最初に性質を見せた後に対象の位置を変化させて、それが同じ対象であるかどうかを判定させる場合、最初に見せた性質と位置替えした対象の性質が同じである方が、再認速度が短くなる(Kahneman et al. (1992))や多重対象追跡実験の結果(被験者に複数の対象のランダムな運動を見せた場合、被験者は4つの対象までは正確に追跡することができる(e.g., Pylyshyn (2007))を説明するためであった。これらの実験結果は、人間が短時間の間は複数の対象の表象を保持する能力を持っていることを示している(Green and Quilty-Dunn (2021), 667)。対象ファイルは、この能力を説明する道具立てとして措定されたものである。

これら二つの特性に加えて対象ファイルがどのような特徴を持つのかについては、論者の間で意見が分かれる。特に、表象媒体として対象ファイルがどのようなフォーマットを持つのかについては、それを図像的(iconic)なものと見なす立場と命題的なものと見なす立場に大別することができる。GreenとQuilty-Dunnは後者の立場をとる。

> [経験的]証拠は、対象ファイルが命題的フォーマットを持つという仮説を支持していると考えられる。命題的表象のように、対象ファイルは個体とその個体が持つ別個な性質の異なる表象から構成される。さらに、命題的表象のように、対象ファイルはそれらの構成要素をより大規模で正確さを評価できる構造にアレンジする。Camp (2007, p. 157) によれば、命題的フォーマットの重要な特徴は、「統語論的構成要素間に成立するある種の機能的関係が、それら構成要素の意味論的値間に成立するある種の論理的あるいは形而上学的関係にマッピングされていること」にある。そうだとすれば、対象ファイルは命題的フォーマットの典型である。(Green and Quilty-Dunn (2021), 677)

GreenとQuilty-Dunnによれば、対象ファイルが表象するのは個体と述語的性質から構成される構造化された内容であり、また、その構造は表象媒体としての対象ファイルの統語論的構造に機能的に反映されている。すなわち、対象ファイルが表象するのは言語的に構造化された命題であり、ファイル自身もそれに対応する統語論的構造を有しているのである。このことは対象ファイルがLOTと極めて類似した道具立てであることを強く示唆している。

ポインター・モデルでは、以上の対象ファイルのアイデアが知覚だけではなく思考などの高次認知状態に拡張して使用される[6)]。対象ファイルで情報が保存されるのはワーキング・メモリーであった。それに対して、ポインター・モデルで情報が保存されるのは長期記憶である。ポインター自身は構造を持たない原子的な表象であるが、情報が保存された長期記憶の場所を指示する働きをする。前節で述べたように、長期記憶に保存された情報は概念(concepts)と区別して「conceptions」と呼ばれる。このconceptionsはプロトタイプや理論など様々な要素から構成されうるが、それらは記述的(descriptive)なものである(Murez (2021))。ここで重要なのは、対象ファイルと同様に、ポインター・モデルもLOTと極めて近い構造を有している点である。ポインター自身は原子的表象として、

6) Quilty-Dunn自身、ポインターのアイデアを、対象ファイルを拡張した心的ファイル理論から借りていることを明言している(Quilty-Dunn (2021), 174)。ただし、心的ファイル理論では概念は心的ファイルに対するラベルと見なされるが、心的ファイルでは多義的な概念をうまく処理できないので、Quilty-Dunnの理論では、概念はあくまでも「記憶の場所」に対するポインターである。

様々な思考の構成要素として登場する。例えば、「トラ」という概念は、トラのプロトタイプなど様々なconceptionsが保存されている記憶の場所を指示するが、それ自体は「トラは肉食である」や「トラはネコ科である」など様々な思考の構成要素となる。ポインターはconceptionsと統語論的に組み合わさることで構造化された命題を表象するのであり、これはLOTの一形態として理解できる。

ポインター・モデルが持つこのような言語的性格は、それを用いたSheaのエピソードDNNが結局のところLOTを密輸入していることを示唆している。たしかに、Millièreが主張していたように、特定内容的計算と非特定内容的計算の違いは程度問題であるという点で、エピソードDNNは古典的計算主義とは異なるメカニズムを有している。しかし、そのことはエピソードDNNが思考の体系性を古典的計算主義とはまったく異なる手段で補っていることを含意するわけではない。思考の体系性の説明を担うのは非特定内容的計算であるが、その実質的説明にLOT的な統語論的構造が前提されているならば、結局エピソードDNNも「体系性のジレンマ」の第二の角を免れることができていないと診断されよう。

では、以上の診断が正しい場合、エピソードDNNに思考の体系性を十分に説明する道は残されているのであろうか。ここで、体系性のジレンマの第二の角に陥らない仕方でエピソードDNNが思考の体系性を説明する方法の一つは、「唯一のゲーム論証」の前提4、すなわち、表象媒体の統語論的構造を説明する唯一の方法はLOTであるという主張を否定することだと考えられるかもしれない。実際、前節で見たように、SheaはDNNやエピソードDNN自体にある種の合成性があることを認めていた。たしかに、これらの合成性はLOTの合成性よりも乏しいと評価されていたが、もしそれが人間の概念的思考をある程度満足の行く仕方で説明できれば、思考の体系性の説明としては十分なものと言えるのかもしれない。

だが、この見込みも極めて低そうである。ここで参考となるのは、認知地図理論である。認知地図理論は1940年代に心理学者のE. C. Tolmanがラットの空間ナビゲーションに関して提案した認知モデルであり、現在でもCampやM. Rescorlaなどが動物や人間の信念の説明として展開している（Camp（2007）; Rescorla（2009））。この理論によれば、少なくとも部分的には、思考の表象媒体は言語的ではなく地図的（cartographic）な統語論的構造を有している。地図的表象は媒体と内容の空間的同型性を有している（Camp（2007）, 159）。地図記号を考えれば分かるように、地図においては、実際の対象は言語以外の記号で表される。これらの記号は現実の対象と似ていなくても構わない。ある対象が別の対象と現実に有している空間的関係が地図において再現されていれば、その地図は外的世界を正しく表象していることになり、不正確に再現していれば誤表象となる。地図的表象は部分を変化させると地図全体が変化するという全体論的性格を有している点で原子論的な合成性を持つLOTとは異なるが、それでも要素に分解・再構成して新しい地図を作ることが可能であり、一定程度の体系性を有している。

SheaがDNNに合成性を認めるときに念頭に置いているのは、認知的地図理論に類したものかもしれない。Sheaの「例えば、それら[DNN]はイメージから対象や性質を抽出することができる」（Shea（2021）, 171）という文言は、地図的表象における統語論的操作を想起させるものである。また、実際Braddon-Mitchell and Jackson（1996）は、地図的な表象理論はコネクショニズムで説明できると主張している。だが、仮にDNNが有する統語論

的構造が言語的なものではなく地図的なものであったとしても、残念ながらそれは思考の体系性の十分な説明とはならない。というのも、CampやRescorlaが認めているように、認知地図理論で説明できる思考の体系性には限界があるからである。空間ナビゲーションなどに使用される比較的単純な思考は地図的表象で十分に説明できるが、抽象的な思考内容や再帰的構造を持つ思考内容を空間的道具立てで説明することは極めて困難である（Camp（2007）, 166-169）。その点で、第3節で述べた意味で、認知地図理論で説明できる概念所有や思考の体系性は、程度として低いものである。この道具立てを用いる限り、LOTを用いた古典的計算主義に比べてDNNの説明力は低い、すなわち、体系性のジレンマの第一の角は免れないと結論付けられることになる。

VII 結論

本稿の議論から結論付けられるのは、エピソードDNNは概念所有を十分な仕方では説明できない見込みが高いという主張である。概念所有の条件の一つに思考の体系性があり、エピソードDNNではその体系性は非特定内容的計算によって説明が試みられる。しかし、非特定内容的計算を概念的思考に適用する際に、Sheaはその概念的思考の説明として概念のポインター・モデルに訴えたが、ポインター・モデルは言語的な統語論的構造を持つ表象媒体を想定するものであり、LOTの一種にコミットするものであった。それゆえ、エピソードDNNによる思考の体系性の説明にはLOTが密かに使用されている見込みが高く、Millièreの思惑に反して、コネクショニズムに従来指摘されてきた「体系性のジレンマ」は十分に克服されていないと考えられる。

もちろん、本稿はエピソードDNNに対するノックダウン・アーギュメントを提供するものではない。Sheaは非特定内容的計算を説明するためにポインター・モデルを用いたが、それ以外のモデルで非特定内容的計算を説明する可能性は残っており、その中にはLOTを前提しないものがあるかもしれない。また、本稿はDNNに対するノックダウン・アーギュメントを提供するものでもない。DNNが思考の体系性を獲得できるという議論には、本稿で扱ったエピソードDNN以外にも、メタ学習に基づく議論やDNNが合成性を獲得できることを論じる別の議論（Lepori, Serre, and Pavlick（2024））があり、その詳細な検討については今後の課題としたい。

【付記】本研究は、科研費22K00030の助成を受けたものである。本稿は、名古屋哲学フォーラム2025～デジタル・ヒューマニティーズと哲学～（2025/02/22　名古屋大学）における同タイトルの口頭発表と、University of WarwickでS. Butterfil教授が主催するResearch Seminarにおける口頭発表（“Deep Neural Networks and systematicity of thoughts,” 2026/02/25）に基づいている。これらの発表でコメントをくださった方に感謝申し上げる。

参考文献

⊙Botvinick, M., Ritter, S., Wang, J. X., Kurth-Nelson, Z., Blundell, C. and Hassabis, D.（2019）. Reinforcement learning, fast and slow. *Trends in Cognitive Sciences* 23（5）: 408-422.

⊙Braddon-Mitchel, D. and Jackson, F.（1996）*The Philosophy of Mind and Cognition: An Introduction*. Malden, MA: Blackwell Publishers.

⊙Buckner, C.（2019）Deep learning: A philosophical introduction. *Philosophy Compass* 14（10）, e12625.

⊙Camp, E. (2007) Thinking with maps. *Philosophical Perspectives* 21: 145-182.

⊙Camp, E. (2009) Putting thoughts to work: Concepts, systematicity, and stimulus-independence. *Philosophy and Phenomenological Research* 78 (2): 275-311.

⊙Coelho Mollo, D. and Vernazzani, A. (ms) Frames of Discovery and the Formats of Cognitive Representation. In Piccinini, G. (ed.) (forth.) *Neurocognitive Foundations of Mind*. Routledge.

⊙Crane, T. (1992) The nonconceptual content of experience. In Crane, T. (ed.) *The Contents of Experience* (Cambridge: Cambridge University Press) : 136-157.

⊙Davis, M. (1991) Concepts, connectionism, and the language of thought. In Ramsey, W., Stich, S. and Rumelhart, D. (eds.) *Philosophy and Connectionist Theory* (Hillsdale NJ: Lawrence Erlbaum): 485-503.

⊙Eliasmith, C. (2013). *How to Build a Brain: A Neural Architecture for Biological Cognition*. Oxford: Oxford University Press.

⊙Evans, G. (1982) *The Varieties of Reference*. McDowell, J. (ed.), Oxford: Clarendon Press.

⊙Fodor, J. (1998) *Concepts: Where Cognitive Science Went Wrong*. Oxford: Clarendon.

⊙Fodor, J. A., and Pylyshyn, Z. W. (1988) Connectionism and cognitive architecture: A critical analysis. *Cognition* 28 (1): 3-71.

⊙Green, E. J. (2018) What do object files pick out?' *Philosophy of Science* 85 (2): 177-200.

⊙Green, E. J. and Quilty-Dunn, J. (2021). What is an object file? *British Journal for the Philosophy of Science* 72 (3): 665-699.

⊙Hummel, J. E. and Holyoak, K. J. (2003). A symbolic-connectionist theory of relational inference and generalization. *Psychological Review* 110 (2): 220.

⊙Hummel, J. E. and Holyoak, K. J. (2005). Relational reasoning in a neurally plausible cognitive architecture: An overview of the LISA project. *Current Directions in Psychological Science* 14 (3): 153-157.

⊙Kahneman, D., Treisman, A. and Gibbs, B. J. (1992) The reviewing of object files: Object-specific integration of information. *Cognitive Psychology* 00324: 175-219.

⊙黒柳奨・柴田正良・戸田山和久・服部裕幸(2003)「コネクショニズムという考え方」、戸田山和久・服部裕幸・柴田正良・美濃正(編)『心の科学と哲学：コネクショニズムの可能性』(東京：昭和堂): 1-25.

⊙Lepori, M., Serre, T. and Pavlick, E. (2024) Break it down: Evidence for structural compositionality in neural networks. *Advances in Neural Information Processing Systems*.

⊙Machery, E. (2009) *Doing without Concepts*. New York, NY: Oxford University Press.

⊙Millière, R. (2024) Philosophy of cognitive science in the age of deep learning. *Wiley Interdisciplinary Reviews: Cognitive Science* 15 (5): 1-9.

⊙Millière, R. and Buckner, C. (2024) *A Philosophical Introduction to Language Models—Part I: Continuity with Classic Debates*. arXiv: 2401.03910.

⊙Murez, M. (2021) Belief fragments and mental files. In Borgoni, C., Kindermann, D. and Onofri, A, (eds.) *The Fragmented Mind* (Oxford: Oxford University Press) : 251-277.

⊙Murez, M. and Recanati, F. (2016) Mental files: An introduction. *Review of Philosophy and Psychology* 7: 265-81.

⊙Pylyshyn, Z. W. (2007) *Things and Places: How the Mind Connects with the World*. Cambridge, MA: MIT Press.

⊙Quilty - Dunn, J. (2021). Polysemy and thought: Toward a generative theory of concepts. *Mind & Language* 36 (1): 158-185.

⊙Quilty-Dunn, J., Porot, N., and Mandelbaum, E. (2023). The best game in town: The re-emergence of the language of thought hypothesis across the cognitive sciences. *Behavioral and Brain Sciences* 46: 1-21.

⊙Recanati, F. (2012) *Mental Files*. Oxford: Oxford University Press.

⊙Rescorla, M. (2009) Cognitive maps and the language of thought. *British Journal for the Philosophy of Science* 60 (2): 377-407.

⊙Shea, N. (2021) Moving beyond content-specific computation in artificial neural networks. *Mind & Language* 38 (1): 156-177.

⊙谷中瞳・峯島宏次(2024)「AIは言語の基盤を獲得するか: 推論の体系性の観点から」『認知科学』31 (1): 27-45.

Deep Neural Networks and Concept Possession

A Critical Assessment of the Episodic Deep Reinforcement Learning Model

Seishu Nishimura

Connectionism is the approach in cognitive science to explain the human mind by means of artificial neural networks. One of the core features of this approach is to realize mental states as distributed representations across many units, which lack any explicit syntactical structure. Between 1980 and 1995, connectionism was criticized for failing to explain the systematicity of thought and concept possession, both of which require the abilities to decompose representations into their parts and to recombine them. Since 2010s, however, deep neural networks (DNNs), a new version of connectionism, have appeared. Raphaël Millière argues for the possibility that DNNs explain the systematicity of thought without using the syntactically structured representations in classical computationalism. To this end, he appeals to episodic deep reinforcement learning (episodic deep RL), which enables DNNs to do both content-specific and non-content-specific computations.

My aim in this paper is to argue that, pace Millière, the episodic deep RL model cannot adequately explain the systematicity of thoughts without smuggling syntactically structured representations. This also shows that the episodic deep RL model cannot treat concept possession by itself. To this end, I will examine Nicholas Shea's arguments about episodic deep RL, which Millière addresses, and argue why Shea's arguments do not fully support Millière.