連合学習理論は擬鼠主義の産物か —表現論としての連合理論—

澤 幸祐¹⁾

Is associative learning theory mere “ratomorphism”?

KOSUKE SAWA¹⁾

Abstract Many researchers have been interested in cognitive abilities of animals and investigated those issues by using several methods. Associative learning theory, which is mainly adopted in classical and operant conditioning literature, has been considered “simple” and sometime “appropriate” account of animal behavior from the point of view of Morgan’s canon. In present article, criticism toward associative account and another reason why associative account is appropriate for animal behavior were reviewed, proposing proper usage of associative learning theory in complex animal behavior. One possible answer in present discussion is that associative learning theory is some kind of “intermediate language”, which can work as the bridge among neural mechanism, psychological faculty in animals, and anthropomorphic explanations.

Key words : associative learning theory, anthropomorphism, ratomorphism

序

1998年の動物心理学会において「擬人主義の功罪」と題したシンポジウムが開催されて14年が経過した。動物行動のメカニズムを考える際に、動物をヒトに擬して理解しようとする擬人主義（anthropomorphism）が取り上げられた理由は、主に比較認知科学という研究分野において、それまでの比較心理学とは異なった問題設定や理念、説明装置が動物心理学会内において一定の広がりを見せ、そこに擬人主義的なものを感じ取った研究者がいたからかもしれない。その後も比較認知科学は国内外において方法論的にも扱う現象に関しても大きな発展を遂げ、動物が示す複雑で豊富な認知的機能の理解について大きな進歩が得られた。翻って、長い歴史を持つ連合学習理論に関する研究動向はどうだろ

うか。筆者の個人的印象では、Rescorla-Wagner model (Rescorla & Wagner, 1972) によって古典的条件づけを巡る理論的研究が大きな転換を経験して以降、様々な理論が提唱されてきたものの（e.g., McLaren & Mackintosh, 2000 ; 2002 ; Miller & Matzel, 1988）、動物心理学全体において占める役割は決して大きなものではなくなってしまった感がある。しかし、隣接領域である神経科学において、ドパミン神経系の挙動が連合学習において古くから指摘されている誤差信号の観点から解釈されるという知見（e.g., Schultz & Dickinson, 2000）が注目を集めるなど、連合学習理論の果たす学術的役割は決して小さくないと思われる。それでもなお、少なくとも本邦の動物心理学分野において、連合学習研究は誤解を恐れず言ってしまうと“不人気分野”で

1) 専修大学人間科学部心理学科

〒214-8580 神奈川県川崎市多摩区東三田2-1-1

連絡先：澤 幸祐

E-mail: sawa@psy.senshu-u.ac.jp

1) Department of Psychology, Senshu University
Corresponding author: KOSUKE SAWA

Published online in J-STAGE: June 21, 2012

doi: 10.2502/janip.62.1.4

ある。連合学習そのものに対する学問的関心は、こと心理学界内においては決して大きくないことは、近年の学会発表動向などを見ても理解できよう。

古典的条件づけという手続きあるいは現象は、ほぼすべての心理学の教科書に登場するほどに文字通り古典的な話題である。しかしながら、多くの心理学者あるいは心理学を学ぶ学生にとっては、「教科書に書いてあることは知っているが、それ以上のことは今更聞けない話題」になりつつあるように思われる。古典的条件づけに関する理論的研究に関わってきた筆者からすれば、こうした状況に対していろいろと思うところはあるが、これは決して現在に限った事でも、日本に限定的な状況ということでもない。例えば、古典的条件づけ理論中興の祖ともいえる Rescorla は、“Pavlovian Conditioning; It's not what you think it is”と題した論文を発表しているし (Rescorla, 1988), 2011年になって出版された連合学習理論とその応用に関する書物の第1章は、“Things you always wanted to know about conditioning but were afraid to ask”と題されている (Schachtman & Reilly, 2011)。すでに20年以上前から現在に至るまで、そして海外においても、連合学習に関しては、今更聞けないという理由からか教科書のかつ表面的な理解が流布していることがうかがえる。

このような状況を受けて、本論文では「古典的条件づけ研究をはじめとする連合学習研究はなぜ不人気なのか」という問題を背景に、擬人主義と対置される「擬動物主義」、あるいは「擬鼠主義」といった観点から連合学習理論の意義を考察する。動物における連合学習に関する研究は、その実験的端緒である Pavlov や Thorndike がイヌやネコを用いていたこと (Pavlov, 1927; Thorndike, 1911) にはじまり、極めて多様な種において検討がなされてきた。それでもなお、動物における連合学習に関しては「ラットを用いた研究から得られた知見」という印象が強いようである。ラット自体は、飼育管理の容易さや多くの方法論的蓄積もあって、確かに広く用いられている動物種であることは間違いない。しかしながら、ラットという種をもって「哺乳類の代表選手」とすることには多くの反論がありうるし、また系統発生的にも適応環境の点からも大きな違いのあるラットから得られた知見を、ヒトにも適用可能な一般原理として安易に捉えることに対する批判も、極めて正当なものである。究極的には「ヒトとは何か」を知りたいのだとすれば、ヒトを直接研究することが近道であろうし (澤, 2008), 動物の中でもラットのような種を用いることに一定の違和感がある人がいてもなんら不思議ではない。「擬鼠主義」

という言葉は、あくまでも「擬動物主義」の持つ否定的な意味を強くしただけの用語であろうが、これは「擬機械主義」と置き換えても類似した議論ができよう。人間は、自らがそれ以外のものによって記述されると何かしらの違和感を生じる。その一方で、心理学者は観察しえない概念を不用意に用いることに強い警戒感を示す。こと動物の「こころ」に関するかぎり、行動の制御に関してこうした概念を用いることにアニミズムの匂いを感じ、抵抗感を持つ向きは多い。Hebb はこれに対し、「行動の事実がそれを避けがたいものにしてしまうと、なぜそれがアニミズム的なのか? 実際は、不可欠な概念を理解可能にする方法を見つけることの中に、本来の苦労があるはずである」と述べている (Hebb, 1949)。本稿では、連合学習による説明を用いて「不可欠な概念を理解可能にする方法」が可能であるかどうかについても考えてみたい。

連合学習理論とは何か

John Locke や David Hume などの英国経験主義哲学に端を発する観念連合論は、Pavlov による古典的条件づけの発見 (Pavlov, 1927), および Thorndike による道具的条件づけ手続きの開発 (Thorndike, 1911) によって、ヒトや動物が示す学習行動の実証データの解釈にも広く用いられるようになった。特に古典的条件づけ研究において、条件刺激 (conditioned stimulus; CS) と無条件刺激 (unconditioned stimulus; US) の対呈示によって生活体が獲得する条件反応 (conditioned response; CR) は、CS と US との間に形成される連合あるいは CS と反応との間に形成される連合によって生じると説明されてきた。すなわち、ベルの音とエサの対呈示試行を経験することによって、ベルの音のみで唾液を流すようになったイヌの例では、「ベルの音とエサの間に興奮の連合が形成された」、あるいは「ベルの音と唾液分泌反応の間に興奮の連合が形成された」と解釈される。動物における連合学習研究の主な目的のひとつは、どのような条件のもとでどのような種類の連合が、どの程度の強度で獲得されるのかを明らかにすることによって、経験によって生じる比較的永続的な行動の変化の原因を記述することにあると要約することができよう。伝統的な意味での連合学習理論とは、こうした連合が形成される際のルール、あるいは刺激の持つ連合可能性 (associability) をはじめとする連合形成に関わる各種パラメータの設定や更新のルールの違いによって差異化されており、上記のような根本的な発想自体は共有されているケースが多い。例えば Rescorla-Wagner model では、CS の明瞭度を示

すパラメータである α は、CSの物理強度によって規定される値であり、経験によって変化することはない。一方、Mackintoshの注意理論(Mackintosh, 1975)やPearce-Hall model(Pearce & Hall, 1980)では、 α はCSに対する注意量、あるいは連合可能性であると仮定されている。CSの物理強度は α の初期値として機能はするものの、 α そのものは経験によって変化するとされている。このように、これらのモデルは α というパラメータの扱いに関して大きく異なるが、こと連合強度の更新に関してはRescorla-Wagner modelとMackintoshの注意理論はBush & Mosteller (1951)の誤差修正に関する枠組みを援用するなどの類似点がある。

ここで、二次条件づけ(second-order conditioning)を例にとって、連合学習理論による説明の一例を見てみよう(e.g., Rizley & Rescorla, 1972; またヒトにおける因果推論と二次条件づけの関連についてはJara, Vila, & Maldonado, 2006も参照)。二次条件づけは、Pavlovによって発見された現象の一つであり、ある刺激AをUSと対呈示した後に別の刺激XをAと対呈示することにより、USとの対呈示を経験していないはずのXに対してもCRが確認されるという現象である(表1)。伝統的な連合学習の作法に基づいた説明のスキーマを図1に示す。図1の左パネルは、Phase 1において行われる手続きによって形成される連合を示している。接近の法則及び随伴性理論に則り、刺激AとUSとの間には興奮の連合が形成されると解釈され、矢印で刺激間に連合が形成されたことが記述される。図1の右パネルには、Phase 2において形成されると予想される連合が示されている。Phase 2においては、刺激Xと刺激Aが対呈示されることにより、これらの間には興奮の連合が形成されると考えられる。その一方で、刺激AはすでにUSとの間に興奮の連合を形成しているためにUS表象を活性化することが可能であると解釈される。また、Phase 1の試行数やUSの種類にある程度依存するものの、刺激Aはある程度のCRを喚起することがあり得る。したがって、刺激XとAの対呈示手続きによって形成される連合はX-A間に限らず、X-US連合やX-CR連合も考慮されなければならない。最終的なテストに

Figure 1. Schematic illustrations of second-order conditioning. Left panel; Associative link between stimuli acquired in Phase 1. Right panel; Associative links acquired in Phase 1 and 2 (solid lines) and those which can be responsible for conditioned responding during test phase (dashed lines).

において刺激Xを被験体に対して提示し、CRが確認された場合には、これらの連合の中でどれが重要な役割を果たしているのかを同定することが必要となる。図1の右パネルに示した中で実線の矢印は、「実験手続き上で実際に対呈示された刺激間の連合、あるいは観察可能な反応との連合」を示し、点線は「刺激XがCRを喚起する上で重要な役割を果たした可能性のある連合経路」を示す。なんらかの実験的処置を加えることで、これらの点線のうちでどの経路が重要な役割を果たしているのかが明らかになれば、「二次条件づけ手続きにおいて刺激XがCRを喚起するメカニズムについて連合学習の立場から説明できた」と考える。一般的には、テストにおいて提示される刺激Xから興奮の連合の矢印をたどってUS表象に、あるいはCRにたどり着くことができ、そうした経路を消去手続きなどによって操作した場合に反応が観察されなくなることなどが確認されることを持って「説明できた」とみなすことが多い。

このように、伝統的な連合学習における説明とは、基本的に機械論的かつ還元論的であるといえよう。

Table 1. Design table of second-order conditioning (SOC).

Group	Phase 1	Phase 2	Test	Result
SOC	A+	XA-	X?	CR

Note; A, X: conditioned stimulus (e.g., light, tone, etc.), +: unconditioned stimulus (e.g., food, electric shock, etc.)

提示された刺激間に連合が形成されるかどうかについては時間的あるいは空間的な接近性、あるいは随伴関係といった外部環境内の変数が重要な役割を果たす。実際に刺激が提示されると、その刺激が連合している様々な刺激表象が興奮・制止といった連合の形式および連合強度に基づいて機械的に活性化され、新たな連合を形成する契機となり、また反応表象が活性化された場合にはCRとして表出される。こうした学習のプロセスは、「どのような連合が形成されるか」あるいは「連合強度の亢進はどのように行われるか」といった側面において違いはみられるものの、基本的には多くの連合学習理論家が共有している思考であろう。

連合学習理論はよい説明か

前項で述べたような基本的な作法を持つ連合学習による説明は、動物行動を説明する上で「よい説明」なのだろうか。確かに、行動主義や新行動主義といった主張が中心であった時代には連合学習による説明はそれなりに受け入れられていた。しかしながら、最初に述べたように、もはや連合学習理論はごく一部の研究者が扱うだけの状況である。本当に「よい説明」であれば、もっと多くの研究者が関心を持ってもよいはずであろう。

連合学習による行動の説明は、しばしば「簡潔で節約的な説明」とみなされることがある。それに伴って、「よい説明」とみなされる場合もある。果たしてそうだろうか。ここで言う「簡潔で節約的な説明」とは、いわゆるモーガンの公準 (Morgan's canon ; Morgan, 1894) に見られるような「動物行動を説明する際には、より低次な能力で説明可能な場合には高次な能力を用いて説明してはならない」という警句に基づいている。賢馬ハンスの例にあるように、動物研究者はこの警句を伝統的に重要視し、自らが得たデータの解釈に細心の注意を払ってきた。そのうえで、連合学習によって説明可能な現象については、推論や思考といった用語を使うことなく動物行動に説明を与えてきたわけである。

ではなぜ、連合学習による説明は「簡潔で節約的」とみなされるのだろうか。確かに連合学習は、ヒトのみでなく多くの動物において確認されるため、進化のプロセスの中では比較的初期に獲得された能力であると推測することができる。また、「刺激と反応の連合」あるいは「刺激と刺激の連合」による説明自体は、生活体に複雑な認知機能を要求することはない。前項で述べたように、連合学習理論は本質的に機械論的な説明を好む。連合学習の作法に則って説明することができるということは、その行動を説明するためには「意識」や「こころ」といったも

のを想定する必要がない、ということを示しており、これを単純であると考えてるのであればまさしく単純な説明であろう。

しかしながら、次のような反論も人口に膾炙している。すなわち、「たった一つの“高次機能”を仮定することと、“百通りの連合経路”を仮定することはどちらが節約的で単純な説明であろうか？」という反論である。この問いに対して、「明らかに後者である」と答えることができて万人が納得するのであれば、擬人主義をめぐる議論など発生しないと思われる。こうした問題は、心の哲学などとも関連して大変に複雑な奥行きを持つものであり、軽々に結論などを出すことは難しい。いずれにせよ、連合学習理論が進歩するにつれて、その理論的背景は格段に複雑になった。前項で述べた二次条件づけの説明は1972年のものであり、それからすでに40年が過ぎようとしている。文脈 (context) 研究や機会設定子 (occasion setting) に関する研究が進み、強化学習理論などとの連携が進む中、もはや連合学習はかつて主張されたような意味においては「単純な説明」ではない。多くの生活体に共有されている機能であり、その基本要素 (要素間の単一連合) が単純なだけであって、連合によって形成されるネットワークそのもの、総体としての連合による説明自体が単純なわけではない。すなわち、連合学習がよい説明であるかどうかの基準として、その「単純さ」を理由とするのは難しいのではないと思われる。

単純さ以外の基準として、神経科学などハードウェアとの対応可能性を候補として考えることができる。ヒトや動物の認知機能のメカニズムを明らかにしていく上で、神経科学が扱う物質的基盤は無視できないものであることは多くの研究者が認めるであろう。もともと、連合学習による説明はその還元論的・機械論的な性質からも神経科学的知見とは相性がよい。例えば Schultz によるドパミン神経系の誤差信号仮説は、Rescorla-Wagner model などの誤差修正過程を学習の基盤に据えるモデルと関連が深い (Schultz & Dickinson, 2000)。同様に、隠蔽 (overshadowing ; Pavlov, 1927) や阻止 (blocking ; Kamin, 1968) のように複数の刺激が存在する中でどのような競合が生じるかに関する議論においても、連合学習理論との整合性が主張されている (e.g., Corlett, Aitken, Dickinson, Shanks, Honey, Honey, Robbins, Bullmore, & Fletcher, 2004 ; Waelti, Dickinson & Schultz, 2001)。こうした知見が連合学習理論によって行動を説明することと与えている妥当性は、「連合学習理論が仮定するプロセスは、生活体の中で実装されている」という点にあると考えられる。行動を説明する理論の価値を

どこに置くかというのは難しい問題ではあるが、実装可能性は大きな魅力であることは否定できない。ヒトや動物の行動を説明するための理論を実現するために、生活体为实现可能な範囲を超えたハードウェアを要求するというのは、明らかに筋違いであろう。もちろん、工学的応用を考えるのであれば生活体が実際に用いているものとは違うが効率のよいアルゴリズムを考案することは高い価値がある。しかしながら、そうした説明スタイルはヒトや動物の行動とその機構に関心のある心理学者の興味とは異なるものであろう。

では、連合学習理論は神経科学のバックアップによって十分に妥当性を担保されるのだろうか。先に述べたように、学習心理学の枠内で扱われてきた連合理論は、実験手続き内で使用されている刺激や観察される反応を「連合する要素」として扱い、これらの間に矢印を結ぶことで構成されるネットワーク構造をもって様々な経験効果を説明してきた。すなわち、連合するものは「刺激」あるいは「刺激表象」、そして「反応」である。したがって、神経科学的知見と連合学習理論の間で十分な対応を取るためには、連合学習理論が仮定する表象が単一あるいは複数の神経細胞によって符号化されており、これらの間に形成されるネットワークが行動の機構として機能することが必要となる。こうした条件自体は、おそらく間違っていないと思われる。神経細胞はお互いに興奮性あるいは抑制性のシナプスを形成してネットワークを作り、その結果として生活体の認知機能を実現しているため、連合学習理論が仮定する構造の多くが実現可能であろう。これは、連合学習理論が仮定する表象間のネットワーク構造が神経細胞間のネットワーク構造と類比的に捉えられるのではないかという希望的観測につながる。しかしながらその一方で、神経細胞の形成するネットワークは、連合学習理論が仮定するそれよりもはるかに複雑な点に注意しなければならない。いわば、連合学習理論は神経系の挙動を過剰に単純化したものであり、連合学習理論が仮定するものが神経系で実現できること自体は決して不思議なことではなく、むしろ当然である。連合学習理論が仮定している連合するものは、先に述べたように操作的に定義可能な刺激表象などであるが、神経ネットワークの単位である神経細胞そのものは、実験心理学において操作的に定義可能である表象である必然性はない。モデルとしてのニューラルネットワークにおいてさえ、ある計算が実現されているネットワーク内の単一ノードが心理学的に意味のある処理や表象に対応しているかと言えば、決してそうではない。言ってみれば連合学習理論は神経系の挙動を我々に理解できる形式に表

現しているだけであり、実際にそうした処理を神経系が実現しているのかどうかは問題のレベルが別である。操作的に定義可能な概念であれば、例えば我々の内省的言語に基づくような連合学習以外の概念であっても神経基盤と対応づけることは可能であり、神経科学的研究において連合学習理論が仮定するプロセスが見出されたところで、それは連合学習理論が他の説明に比べて優れているものと主張する根拠としては決して強いものではない。

表現論としての連合学習理論

ここまで、連合学習がよい説明であるとされる理由についていくつかの話題を取り上げ、そのいずれもが決して連合学習が他の説明スタイルと比較して優れたものであることを強く主張できるものではないことを見てきた。だとすれば、連合学習は「よい説明ではない」ということになるのだろうか。

前項で述べたように、連合学習理論は神経科学的研究と相性がよい。このことが直ちに連合学習理論の優位性を担保するものではないことはすでに述べた。その一方で、神経系の挙動を表現するものとしての連合学習による説明自体は、その価値を減じるものではない。我々は、神経系が行っている処理を理解する上で、連合学習理論が仮定するメカニズムを「表現論」として利用することはできる。もちろん、神経系が行っている処理を擬人的説明やヒトの内省的用語で表現することも可能であり、このことが神経系との対応を根拠に連合学習理論の優位性を主張できない理由である。この時点では、神経系が行っている処理と実際の行動の間をつなぐインターフェースとして、擬人的説明と連合学習理論の間には優位性に関する違いを見出すことができない。そこで、ひとつのアナロジーを考えてみたい。すでに認知科学などにおいては手垢のついた「コンピュータアナロジー」である。

我々がコンピュータを利用する際にはプログラムを書く。プログラミング言語には多くの種類があり、機械語のようにコンピュータがそのまま実行できるようなものもあれば、C言語やPython, Javaといった人間が理解しやすい文法をもった言語も存在する。人間が理解しやすいように構築されたプログラミング言語は高級言語と呼ばれ、そのままでは機械が実行できない。そのため、高級言語で書かれたプログラムは、コンパイラを用いて実行形式に変換するかインタプリタと呼ばれるプログラムを用いて逐次実行するなどの方法を取って初めてコンピュータで実行することができる。我々が実験機器の制御などに用いるのは多くの場合高級言語であり、機械自体が行うべき処理を別形式で、「ヒトが理解しやすいよ

うに」表現しなおしたものであると考えることができよう。実際には、多くのコンパイラは、高級言語で書かれたソースコードを機械語へと直接翻訳するのではなく、中間言語と呼ばれる別の表現に翻訳し、最適化した上で機械語へと変換する。つまり、「実際のハードウェアの処理」と「ヒトが理解しやすいような言語」との間には、これを媒介する中間表現が存在する。

こうした構造は、生活体内の神経ネットワークに組み込まれているノードや要素の処理が、我々の持つ言語体系と直接的に対応しているが不明でありながらも、それが実現している処理は「擬人的説明」によって解釈できるという問題と類似しているように思われる。すなわち、「擬人的説明」あるいは「ヒトの内省的言語による説明」とは、いわば我々のコンピュータのCPUが行っている処理を高級言語で記述されたプログラムで表現しているような状況と考えることができる。では、連合学習理論による説明とはどのように理解することができるだろうか。連合学習理論もまた、神経系が行っている処理の表現としてとらえることができる程度可能であるが、その文法は我々が普段用いている「内省的言語」とは異なる。すなわち、CPUが行っている処理を何らかの形で表現はしているものの、高級言語が持っているようなわかりやすさを持ち合わせているわけではない。これは先に述べたコンピュータアナロジーからいえば、中間言語のようなものと考えることが可能だと思われる。実際のハードウェアが行っている処理と人間が理解しやすい説明との中間に位置し、これらのインターフェースとして機能するような中間言語として連合学習理論はその意義を持ちうるのではないだろうか。

連合学習理論を中間言語としてとらえることの有用性は、本稿の中心の問題でもある「擬人主義」と「擬鼠主義」の対置に関して有効なアプローチになりうるという点にある。擬人的説明とは「ヒト以外の動物が示す行動をヒトに擬して説明すること」であり、一方の擬鼠的説明とは「ヒト（あるいはラット以外の動物）が示す行動をラットに擬して説明すること」である。これは、ヒトとラットを交換しただけで対称な関係にあるようにみえる。しかしながら、実際にはここには正確な対称関係は存在しない。擬人的説明の多くが「ヒトの内省的言語」に基づくのに対して、擬鼠的説明として我々は「ラットの内省的言語」を使っているわけではない。あくまでも、ヒトの内省的言語以外のものを使用しているだけであり、そこには対称的な関係は存在せず、ヒトと動物を対置させているわけでも並立させているわけでもない。また、動物が行っている情報処理の

中に、ヒトの内省的言語では記述できない、あるいは記述が困難なものが含まれていた場合には、正確な理解そのものが阻害されてしまうことになる。これに対して、ヒトの内省的言語とハードウェアの間にある中間言語としての連合学習理論はどのように位置づけられるだろうか。連合学習理論による説明は、ヒトの内省的説明、擬人主義とは相いれない側面を持っているかもしれないが、先に述べたように神経系の挙動の表現としては十分な有用性を持つ。ヒトの内省的説明を用いるよりも、より実際の神経系の挙動を捉えるのに連合学習理論という中間言語を用いることは正確な記述を可能にすると思われる。また、連合学習理論が記述する対象は神経系の挙動に留まらない。ヒトや動物の行動の背景にある至近要因としての認知機能を記述する際にも用いることができる。ヒトの持つ機能や内省的言語を暗黙に前提してしまうような説明ではなく、一度それを連合学習理論という中間言語に翻訳することによって、ヒトと動物を共通する土台に載せて比較し、議論することが可能になると思われる。

中間言語としての連合学習理論は、文字通り「言語」のようなものであり、生活体の認知機能や神経系の挙動を記述するためのものである。言語である以上、そこには「文法」が存在する。文法に従ってさえいけば多種多様な「文章」を作ることができる。ここでいう多種多様な文章、すなわち連合学習理論という言語によって作り出される表象間のネットワークの多様性とは、異なる認知機能を連合学習理論という言語で表現したり、あるいは異なる種が見せる異なる行動や機能を連合学習理論という言語で表現することなどを可能にする。連合学習理論はあくまでも「表現」であり、その先には「表現されている対象」が存在する。動物行動研究において「表現されている対象」の多くは、我々の持つ自然言語体系に基づく内省的言語などでは到底記述できるものではなかろう。そうしたものを、ヒトに固有のものに縛られずに統一的に表現する説明体系としての意義こそが、連合学習理論の持つ中間言語としての特徴であろう。

連合学習理論をめぐる種々の対立

最初に見たように、連合学習理論に基づいた研究は、少なくとも日本国内においては「不人気分野」であり、いわば帰無仮説として扱われてきた (e.g., 井口, 2008)。帰無仮説である以上、そこには対立仮説が存在することになり、様々な点において連合学習理論による説明の妥当性に関する対立構造が存在する。

これまで連合学習理論は「比較的単純な学習効果

を説明するためのもの」と考えられてきた。すなわち、そこには「連合学習とは対立する高次な認知機能」なるものが存在することが含意されている。では、連合学習理論による説明は、「高次認知」と呼ばれるものと排他的に対立するものなのだろうか。これについては様々な議論が存在するが (e.g., Buckner, 2011; Shettleworth, 2010), 前項の議論から考えると、決して排他的に対立するものではないと考えられる。問題はあくまでも「連合学習理論によって記述可能か」であって、排他的に対立するような異なるメカニズムが存在するかどうかではない。連合学習理論はあくまでも記述言語であるので、その中で、既存の連合学習理論の文法では記述不可能な問題が発生した場合には、文法を見直せばよい。説明できない現象が現れた時には理論体系を改訂するというのは、連合学習理論の歴史の中でも繰り返されてきたことである (e.g., Dickinson & Burke, 1996; Van Hamme & Wasserman, 1994)。すなわち、中間言語としての連合学習理論という立場を取る限りにおいて、いわゆる「高次認知」との間に対立構造を設定する必要は本来的にはないと思われる。

連合学習理論は、その名の通り「経験によって生じる比較的永続的な行動の変化」としての学習に関して用いられてきた説明である。したがって、その適用範囲は基本的に経験効果に限定されてきた。動物が何かしらの行動を示した際に、その行動が動物の経験した訓練履歴に依存したものであると考え、そのメカニズムを連合学習理論によって記述するという具合である。その一方、動物行動に関心を持つ研究者の中には、動物が示す行動を生得的な機構によって説明することを目指す人々もいる。その多くは、動物行動の背景にあるメカニズムを進化の過程において獲得されてきたものとして捉えつつ、後天的な訓練によって生態学的妥当性抜きに説明されてしまうような連合学習理論による記述を「帰無仮説」として対立させる傾向にあるように見える。ここでの対立は、動物が示すある行動が生得的なものであるか後天的な経験によるものかという議論と関係するが、これは連合学習理論による説明の妥当性以前の問題であり、初期経験のはく奪などの実験的操作によって検討すればよい。この対立に関する真の問題は、生得的な機構に基づいて発現する行動やその背景にある認知機能に関して連合学習理論の文法を適用してよいか、という点である。これは、生得的な機能に後天的な機能の説明を当てはめるという意味で矛盾した問いのように見えるかもしれない。しかし、次のような例を考えてみよう。我々の経験する味覚は、甘味・塩味・酸味・苦味の4基本味から

なる。これらの味覚刺激は生活体に對し、快や不快の顔面反応を喚起させることが知られているが、ここで味覚刺激が喚起する顔面反応は生得的なものであることが報告されている (e.g., Rosenstein & Oster, 1988)。その一方で、Berridge, Grill and Norgren (1981) は、味覚嫌悪学習によってこうした顔面反応が変化することを報告している。こうした知見は、味覚刺激と顔面反応の間に存在する連合が生得的に存在する一方で、この連合が後天的な経験によって上書き、あるいは変容すると捉えることによって、連合学習理論の文法を用いた統一的説明を与えることができる。生得的に存在する機構に関して、連合学習理論の文法を適用すること自体は決して伝統的な立場と一致するものではないが、統一的な記述を与えるという意味では考慮されるべき拡張であろう。

このように、連合学習理論に関する対立の多くは、連合学習理論のある程度の拡張と、中間言語としての性質を利用することで解消可能なものが数多く存在する。排他的な対象として捉え、帰無仮説として棄却するのではなく、複数の種の類似と相違を含みつつ、神経基盤や言語化されない認知機能と我々の理解をつなぐ中間表現として、連合学習理論を用いた説明は様々な有用性を秘めていると考える。

連合学習理論は、今後どうあるべきか

このように、連合学習理論は、単にラット（あるいはヒト以外の動物全般）から得られた知見をもとにヒトを動物に擬するだけのものではなく、また伝統的に連合学習理論家が仮定していたような「生活体全般に適用可能な統一原理」でもない、ある種の言語体系であるにとらえるべきだと考えられる。すなわち、「ヒトはラットと同じだ」と主張することも、「ラットもヒトのようなものだ」と主張する必要もなく、「ラットとヒトはこういう点で同じで、こういう点で違う」ということを、一方に強く偏ることなく統一的な体系で記述することができる。それは本来、動物心理学や比較心理学が目指したものではなかったか。現状の連合学習理論が、こうした可能性を十分に体现したものであるとは多くの研究者は納得しないであろう。では、連合学習理論は今後どうあるべきなのだろうか。

連合学習理論の持つ魅力であり、また大きな問題でもあるのは、「表象間の結びつきですべてを記述する」ことの単純さと、そこから来る機械論的色彩であろう。連合学習理論が使用する表象とは、あくまでも刺激や反応の代替物としての意味合いが強く、いわば外部環境や実験手続きによって制御されている変数間の関係を表現したものである。そこには、

能動的な“意識”といった要因は介在しにくい。動物を用いた意識研究というものには多くの研究者は慎重な姿勢をとるかもしれないが、魅力的な研究領域であることは間違いない。果たして連合学習理論は「意識」について語ることはできるか。現状においては、明らかにそれは難しいであろう。連合の鎖から能動的な意識が創発するという証拠はもとより、能動性を記述することすら困難である。また、連合学習理論において連合する対象は刺激や反応に対応する表象であるが、果たしてこうした表象は連合の最小単位なのだろうか。すでに議論したように、連合学習理論の仮定する表象のレベルは、US 表象に情動的属性を組み込む理論 (Wagner & Brandon, 1989) などのように精緻化を試みる例や、逆に生活体が経験する環境全体を一つの刺激として捉え、実験者が提示する刺激を連合の単位というよりもむしろ刺激要素とみなす理論 (Pearce, 1987) も存在するが、基本的には単純な形式で記述されることが多い。生活体が経験する環境や刺激をどのように記述し、いかに精密な表象を構成して理論に組み込むかについての議論は、連合学習理論のよりよい拡張に重要な意味をもつだろう。

いずれにせよ、物言わぬ動物の持つ認知機能や神経基盤について何かを知ろうとした時に、連合学習理論のような中間表現があることは有益である。最後に、中枢神経系の活動を計算機科学から論じたフォン・ノイマンの言葉を紹介する (von Neumann, 2000)。「私たちが数学を語るときには、中枢神経系が現に用いている一次言語の上に構築された二次言語について語っているのかもしれない。したがって、中枢神経系が実際に使っている数学的言語あるいは論理的言語が何であるのかを見極めるという観点に立つと、私たちの数学の外形は絶対的な重要性を持たない。私たちのもつ内省的言語、あるいは擬人的説明は、動物の中で何が起きているのか、神経系が何をしているのかを理解する上では、決して絶対的なものではない。しかし、我々はそれを理解したい。中間表現の存在は、その助けになりこそすれ、邪魔になることはないであろう。

謝辞

本論文は、日本動物心理学会第71回大会 (Animal 2011) において開催されたシンポジウム「動物研究最前線：擬人主義とどうつきあうか」での話題提供に基づいたものである。学会およびシンポジウムの開催にご協力いただいた方々に感謝したい。なお本稿の作成にあたっては、平成23-27年度文部科学省私立大学戦略的研究基盤形成支援事業「融合的心理科学の創成：心の連続性を探る」(S1101013) による

助成を受けた。

引用文献

- Berridge, K. C., Grill, H. J., & Norgren, R. 1981 Relation of consummatory responses and preabsorptive insulin releases to palatability and learned taste aversions. *Journal of Comparative and Physiological Psychology*, **95**, 363-382.
- Buckner, C. 2011 Two Approaches to the Distinction between Cognition and 'Mere Association'. *International Journal of Comparative Psychology*, **24**, 314-348.
- Bush, R. R., & Mosteller, F. 1951 A mathematical model for simple learning. *Psychological Review*, **58**, 313-323.
- Corlett, P. R., Aitken, M. R. F., Dickinson, A., Shanks, D. R., Honey, G. D., Honey, R. E., Robbins, T. W., Bullmore, E. T., & Fletcher, P. C. 2004 Prediction error during retrospective revaluation of causal associations in humans: fMRI evidence in favor of an associative model of learning. *Neuron*, **44**, 877-88.
- Dickinson, A., & Burke, J. 1996 Within-compound associations mediate the retrospective revaluation of causality judgments. *Quarterly Journal of Experimental Psychology*, **49B**, 60-80.
- Hebb, D. O. 1949 *The Organization of Behavior: A Neuropsychological Theory*. John Wiley & Sons, Hoboken, NJ. (ヘップ D. O. 鹿取廣人・金城辰夫・鈴木光太郎・鳥居修晃・渡邊正孝 (訳) 2011 行動の機構 脳メカニズムから心理学へ 岩波書店)
- 井口善生 2008 連合論的学习心理学と比較認知科学の断絶と接点 動物心理学研究, **58**, 61-68.
- Jara, E., Vila, J., & Maldonado, A. 2006 Second-order conditioning of human causal learning. *Learning and Motivation*, **37**, 230-246.
- Kamin, L. J. 1968 Attention-like processes in classical conditioning. In M. R. Jones (Ed.), *Miami Symposium on the Prediction of Behavior: Aversive Stimulation* (pp.9-32). Coral Gables, FL: University of Miami Press.
- Mackintosh, N. J. 1975 A theory of attention:

- Variations in the associability of stimuli with reinforcement. *Psychological Review*, **82**, 276-298.
- McLaren, I. P. L. & Mackintosh, N. J. 2000 An elemental model of associative learning: I. Latent inhibition and perceptual learning. *Animal Learning & Behavior*, **28**, 211-246.
- McLaren, I. P. L., & Mackintosh, N. J. 2002 Associative learning and elemental representation: II. Generalization and discrimination. *Animal Learning & Behavior*, **30**, 177-200.
- Miller, R. R., & Matzel, L. D. 1988 The comparator hypothesis: A response rule for the expression of associations. In G. H. Bower (Ed.), *The psychology of learning and motivation* (Vol.22, pp.51-92). San Diego, CA: Academic Press.
- Morgan, C. L. 1894 *An introduction to comparative psychology*. London: Walter Scott.
- Pavlov, I. P. 1927 *Conditioned reflexes*. Oxford: Oxford University Press.
- Pearce, J. M. 1987 A model for stimulus generalization in Pavlovian conditioning. *Journal of Experimental Psychology: Animal Behavior Processes*, **13**, 182-192.
- Pearce, J. M. & Hall, G. 1980 A model for Pavlovian learning: Variations in the effectiveness of conditioned but not of unconditioned stimuli. *Psychological Review*, **87**, 532-552.
- Rescorla, R. A. 1988 Pavlovian conditioning. It's not what you think it is. *The American Psychologist*, **43**, 151-60.
- Rescorla, R. A., & Wagner, A. R. 1972 A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. In A. H. Black & W. F. Prokasy (Eds.), *Classical conditioning II: Current research and theory* (pp.64-99). New York: Appleton Century Crofts.
- Rizley, R. C., & Rescorla, R. A. 1972 Associations in second-order conditioning and sensory preconditioning. *Journal of Comparative and Physiological Psychology*, **81**, 1-11.
- Rosenstein, D., & Oster, H. 1988 Differential facial responses to four basic tastes in newborns. *Child Development*, **59**, 1555-1568.
- 澤幸祐 2008 動物心理学における研究パラダイム間の「翻訳」について 動物心理学研究, **58**, 73-76.
- Schachtman, T. R., & Reilly, S. 2011 *Conditioning and animal learning: Human and Non-human applications*. New York: Oxford University Press.
- Schultz, W. & Dickinson, A. 2000 Neuronal coding of prediction errors. *The Annual Review of Neuroscience*, **23**, 473-500.
- Shettleworth, S. 2010 *Cognition, evolution, and behavior*, 2nd ed. Oxford, UK: Oxford University Press.
- Thorndike, E. L. 1911 *Animal intelligence: Experimental studies*. Macmillan.
- Van Hamme, L. J., & Wasserman, E. A. 1994 Cue competition in causality judgments: The role of nonpresentation of compound stimulus elements. *Learning & Motivation*, **25**, 127-151.
- von Neuman, J. 2000 *The computer and the brain. (second edition)*. Yale University Press. (フォン・ノイマン J. 柴田裕之 (訳) 2011 計算機と脳 ちくま学芸文庫)
- Waelti, P., Dickinson, A., & Schultz, W. 2001 Dopamine responses comply with basic assumptions of formal learning theory. *Nature*, **412**, 43-48.
- Wagner, A. R., & Brandon, S. E. 1989 Evolution of structured connectionist model of Pavlovian conditioning (AESOP). In S. B. Klein and R. R. Mowrer (Eds.), *Contemporary learning theories: Pavlovian conditioning and the status of traditional learning theory*. Hillsdale, NJ: Erlbaum.

(2012. 1. 11 受稿, 2012. 5. 1 受理)