東京大学教養学部哲学・科学史部会　哲学・科学史論叢第二十号　平成 30 年 3 月 (59–84)

誤表象問題
―学習に基づく目的論的意味論の検討―

勝亦　佑磨

1.　序論――目的論的意味論に基づく志向性の自然化と誤表象問題

1.1　志向性の自然化

私たちの心の持つ重要な特徴として,「志向性 (intentionality)」が挙げられる [1]. 志向性とは，あるものがそれ自身とは別の何かを表象するという性質である. 例えば,「目の前に熊がいる」という信念は「目の前に熊がいること」を表象しているとともに,「目の前に熊がいる」いう内容（表象内容）を持つという意味で, 志向性を持つ. とはいえ私たちは, 実際には目の前に熊がいないとき, 例えば目の前にある大きな木を熊だと間違える（誤表象する）ときにもまた,「目の前に熊がいる」という信念を持ちうる. このような誤表象のケースにおける「目の前に熊がいる」という信念もまた,「目の間に熊がいる」という内容を持つという意味で，志向性を持つ. このように，心が志向性を持つこと，すなわち表象することには，常に，誤って表象する可能性が伴っている.

それでは，私たちの心は，なぜ志向性を持つのだろうか. 単に心が心であるがゆえに志向性をもつのだという考えは，パトナム (Putnam, 1981) によって「指示の魔術理論」と呼ばれ, 問題があるとして批判されてきた. そうだとすると, 心がなぜ志向性を持つのかを説明する必要がある. こうした問題に対して，心の持つ志向性を自然化する立場，すなわち自然科学と調和するような一元論の枠組みで志向性を説明しようとする論者たちは，それぞれの仕方で解決を試みてきた.

志向性を自然化する代表的な 2 つの理論として，初期のドレツキが『知識と情報の流れ (Dretske, 1981)』で展開した理論をはじめとする情報意味論と，ミ

リカンをはじめとする目的論的意味論 (Millikan, 1989) が挙げられる．一般に，志向性を説明する理論としては，後者の目的論的意味論の方が有力であるとされている．というのも，ドレツキによる情報意味論は，志向性の重要な側面，すなわち誤表象の可能性をうまく説明できないという点で問題を抱えているためである．

1.2 初期ドレツキによる情報意味論と誤表象問題

それでは，ドレツキによる情報意味論は，どうして誤表象を説明できないのだろうか．端的に述べれば，ドレツキは，表象するものと表象されるものの関係を，厳密な法則による関係，すなわち情報関係によって定義していたためであると考えられる．この点に関して，以下で，詳細に説明しよう．

まず，ドレツキは「情報」[2] を，自然世界の客観的で心から独立の特徴をもつものとして扱う．そして，信号 r が「S が F である」という情報を運ぶのは，信号 r が得られたという条件のもとでの「S が F である」ことの「条件付確率が 1」であるときである．つまり信号 r が得られるときは，「S が F である」ときであり，そのときに限られるということである (p.65)．そのうえで，ドレツキは，表象関係を次のように定義する．「A が B を表象する」のは，A が B によって因果的に引き起こされ，A が B についての「情報を運んでいる」ときであり，そのときに限られる [3]．ここで重要なのは，A が存在するときには必ず B が存在するという点である．例えば，私の「目の前にヘビがいる」という知覚 (A) が「目の前にヘビがいること」(B) を表象するとすれば，「目の前にヘビがいる」という知覚は「目の前にヘビがいること」のみによって因果的に引き起こされるのであり，「目の前にヘビがいる」という知覚があるときには，常に「目の前にヘビがいる」ということになる．

しかしながら，こうしたドレツキ初期の表象論は，「選言問題」に陥ってしまい，誤表象をうまく説明できないために，多くの批判を受けてきた（例えば Fodor, 1984）．例えば，私たちは道に落ちている木の枝をヘビと見間違えると

きがある．このとき，「目の前にヘビがいる」という知覚 (A) は，「目の前に木の枝があること」(B) によって引き起こされている．したがって，ドレツキ初期の表象論では，「目の前にヘビがいる」という知覚 (A) は，「目の前にヘビがいること，または目の前に木の枝があること」(B) を表象することにならざるをえない．そうすると，「目の前にヘビがいる」という知覚は「目の前に木の枝があること」によって引き起こされても，それが誤りであるということはできない．そして，知覚 (A) は「目の前にヘビがいる」という知覚であるというよりも，「目の前にヘビがいる，または目の前に木の枝がある」という知覚だという方が適切である．このように，ドレツキ初期の厳密な法則による情報関係による表象の理論は，選言問題，すなわち表象が選言的な内容を持つために誤表象が不可能になるという問題に陥ってしまい，誤表象を説明できないのである．

1.3 ミリカンによる目的論的意味論

一方，ドレツキによる情報意味論に対して，ミリカンによる目的論的意味論は，こうした誤表象の問題に一定の解決策を与えていると考えられる．まず，目的論的意味論とは，どのような立場なのかを以下で説明しよう．

目的論的意味論とは，心的状態を生物の心臓や肺のように生存に有益な機能を持つ状態として捉えることで，志向性を説明する立場である．目的論的意味論の主流な立場はミリカンによる進化に基づく表象論であり，それは次のようなものである．生物の器官や特質は，それらが持つ機能のおかげで生物の祖先の生存に役立ち，それゆえそれらは進化の過程で選択されてきた．このようにして選択されてきた生物の器官や特質の持つ機能は目的論的機能と呼ばれる．例えば，心臓は，体内に血液を循環させるという機能を持つために祖先の生存に役立ち，それゆえ進化の過程で選択されてきた．それゆえ心臓は，血液を循環させるという目的論的機能を持つと考えられる．

同様に，心的状態の持つ志向性もまた，こうした目的論的機能によって説明

される．心的状態を生み出す形成機構（生産者）とそれに基づいて行動を引き起こす利用機構（消費者）は，それぞれの機能を持つために祖先の生存や繁殖に役立ち，それゆえに進化の過程で選択され，現在においても存続している．例えば，あるタイプの知覚の形成機構と利用機構の組は，「目の前にヘビがいるとき」にその知覚を生み出し，それに基づいて「目の前のヘビから逃げる」という行動を引き起こしたことが祖先の生存に役立ち，そのため，進化の過程で選択されてきた．それゆえ，この知覚の形成機構と利用機構は，そうした目的論的機能を持つということになる．同様に，こうした形成機構によって生み出され利用機構によって行動に用いられるような個々の知覚もまた，「目の前にヘビがいる」ときに「目の前のヘビから逃げる」という派生的な目的論的機能[4]を持つことになり，それゆえ「目の前にヘビがいること」を表象するのだと考えられる．

それでは，ミリカンによる目的論的意味論は，誤表象をどのように説明するのだろうか．ミリカンは，誤表象を，目的論的機能がうまく働かないときに生じるものとして説明する．例えば，知覚の形成機構と利用機構の組が，目の前に木の枝があるときに，「目の前にヘビがいる」という知覚を生み出し，逃げるという行動を引き起こすような場合，知覚の形成機構と利用機構の組の機能はうまく働いておらず，誤表象が生じていると考えられる．誤表象は，このように行動の失敗と結び付けて説明される．目の前に実際にヘビがいる際に逃げるという行動が引き起こされるならば，その行動は生存に貢献するようなものであり，成功していると考えられるが，目の前にヘビがいないにもかかわらず，逃げるという行動が引き起こされるのであれば，その行動は生存に貢献しているとはいえず，失敗であると考えられるのである．

このようにミリカンの理論は，表象がどう生産されるかだけでなくそれが行動においてどう消費（利用）されるのかという両方の側面をふまえた，目的論的機能によって表象内容を説明するという特徴がある．ミリカンの理論が誤表象に一定の説明を与えることができているのは，こうした消費の側面が考慮されているためであろう．こうしたミリカンによる理論と比較すると，先にみた

ドレツキ初期の情報意味論は，表象の生産の側面，すなわち表象するものと表象されるものの間の関係のみを問題にする理論であった．こうした二者の関係のみを問題にするドレツキ理論にしたがえば，先にみたように，表象が選言的な内容を持つことになり，誤表象が不可能になってしまうことになる．その意味で，ミリカンによる目的論的意味論と比較し，ドレツキ初期の情報意味論は，志向性を説明する理論としては欠陥があるといわざるをえないだろう．

1.4 ドレツキによる新たな表象論と本論文の目的

しかし，ドレツキは，後に『行動を説明する』(1988) において，『知識と情報の流れ』で展開した初期の情報意味論を基礎にしつつも，それを大幅に修正し，独自の目的論的意味論を展開した．そこでは，以前のように表象の生産の側面を重視しつつも，消費の側面を取り入れた新たな表象論が展開されており，以前の著作において課題であった誤表象問題に一定の解決策を与えている可能性がある．しかもこの解決策は，ミリカンの目的論的意味論とは異なる方法によるものであると考えられる．というのもドレツキは，ミリカンのような進化に基づく表象論ではなく，学習に基づく表象論を展開しているからである．それにもかかわらず，ドレツキによる学習に基づく表象論が誤表象問題を解決できているかどうかに関する十分な検討は，ほとんど行われてこなかった[5]．本論文では，こうした背景をふまえて，『行動を説明する』における表象論を検討していきたいと思う[6]．

本論文の意義は，志向性の自然化の問題を，目的論的意味論の主流である生物の進化の視点ではなく，あえて学習という視点から解く可能性を検討することにある．とはいえ，ドレツキ自身もそうであるが，本論文でも，進化に基づく理論を否定するつもりはない．本来，志向性は個々の生物の心的状態が持つ性質であるという意味では，すべて個体における志向性である．しかし，心的状態が志向性をもつのはなぜかという問題に対して，進化における種レベルの説明が可能な場合と，学習による個体レベルの説明が可能な場合がある．少な

くともドレツキは，これらの説明が異なる種類のものであると考え，学習による表象論の必要性を訴えている[7]．だが，もしそうだとしても，まずは，この新ドレツキ表象論が，以前の著作から課題であった誤表象問題を解決できているのだということを示さないことには，志向性を説明する理論として受け入れることすらできないであろう．

本論文の構成は，以下の通りである．2 節では，『行動を説明する』で提示されたドレツキの新たな表象論をみて，それがどのようにして誤表象問題への解決を試みているのかを検討する．3 節では，ドレツキ表象論が少なくとも 3 つの問題を抱えており，実際には誤表象問題を解決できていないことを示す．4 節では，こうしたドレツキ表象論がなぜ誤表象問題を説明できていないのかを，より詳細に検討する．そこでは，表象の消費の側面を取り入れたとはいえ，初期の理論と同様，依然として表象の生産の側面を重視した理論であるために問題を抱えているのだということを明確化する．5 節では，結論と，今後の課題を述べる．

2. 『行動を説明する』におけるドレツキ表象論――誤表象問題への解決策

本節では，ドレツキの『行動を説明する (1988)』のなかで修正された新たな表象論をみて，ドレツキがいかにして誤表象問題を解決しようとしたのかを示す．

2.1 ドレツキによる新たな表象の定義

まず，ドレツキによる新たな表象の定義を説明しよう．それによれば，「C が F を表象する」のは，「C が F を表示する (indicate) 機能を持つ」ときであり，そのときに限られる (Dretske, 1988, pp84f)．それでは，「C が F を表示する機能を持つ」とはどのようなことだろうか．それを説明するために，まずは「C が F を表示する」[8] というのはどのようなことなのかを説明しよう．

C が F を表示するのは，C が得られたという条件のもとでの「F であることの条件付確率が 1 である」ときであり，そのときにかぎられる．こうした表示関係は，トークン間の関係であるが，そのトークン間の関係が表示関係であるのは，ある事象タイプとある事象タイプの間に次のような関係が成り立っている場合である．すなわち，C タイプの事象が得られたという条件のもとでの「F タイプの事象があることの条件付確率が 1 である」という関係が成り立っているときかつそのときに限り，C タイプのあるトークンは，F タイプのあるトークンを表示することになる [9]．

ここで重要なのは，表示関係が成り立っているときには必ず，F なしに C は生じないという関係が成り立っているということである．例えば，ある森で，あるタイプの足跡 (C) を残すのがウズラ (F) だけであり，ウズラ以外にはそのタイプの足跡を残すような動物（や他の要素）が存在しないのであれば，その足跡タイプ C のあるトークンはウズラ (F) を表示するといえる．定義上，こうした自然的記号に，「誤表示」はありえない．例えば，別の森では，ウズラだけでなくキジもまた同じタイプの足跡を残すとしよう．この森において，その足跡は，ウズラもキジも表示せず，それらの間に表示関係は成り立たない．

ドレツキによれば，動物のいくつかの内的状態もまた，いくつかの外部状態とそれぞれ表示関係にある．例えば，学習前において，ある動物の内的状態 C（赤の知覚）は外部状態 F（赤であること）と表示関係にあり，C は F の表示子である [10]．ドレツキによれば，こうした単なる表示子である C は学習を経て，「F を表示する機能」を獲得し，表象になる．すなわち，誤表象の可能性が生まれるのである．だが，いかにして学習によって内的状態 C が表示機能を獲得するか（すなわち表象になるのか）を説明する前に，まずはドレツキにとって，表象するとはどのようなことなのかを，以下で説明しよう．

先にも述べたが，ドレツキにとって「C が F を表象する」とは，「C が F を表示する機能を持つ」ことである（このとき，C は F という表象内容を持つ）．それでは，「C が F を表示する機能を持つ」とはどのようなことであるかを説明しよう．「C が F を表示する機能を持つ」という表現において，「C」は「F

を表示する」の主語であると同時に，「F を表示する機能を持つ」の主語でもある．この表現において，前者の場合はトークン間の関係を，後者の場合はタイプ間の関係を意味すると考えられる．

また，「C が F を表示する機能を持つ」という表現は，この機能がうまく働くときもあればそうでないときもあることを含意している．この機能がうまく働くときには，C のあるトークンは（実際に存在している）F のあるトークンを表示する[11]．すなわち，C は F を正しく表象していることになる．一方，この機能がうまく働かないときはどうだろうか．F のトークンが実際には存在していないときにも，C のあるトークンは F という内容を持つ．このとき，C のあるトークンは，本来表示すべきである F のトークンを表示しようとしているが，実際には表示していないために，誤表象していると考えられる．

初期の表象論と比較して，新たな表象論に，ドレツキがこうした機能の概念を導入したのは，次のためであろう．すなわち，「C が F を表象する」ことを「C が F を表示する機能を持つ」ことであると定義し直すことで，C が存在するときに必ずしも F が存在している必要がないという余地を残したという点である．こうした余地を残すことで，誤表象が可能になると，ドレツキは考えたのである[12]．

2.2　行動における学習と、表示機能の獲得

先にも述べたが，行動[13]における「学習」を経て，生物の内的状態 C（例えば脳状態）は「F を表示する機能」を獲得し[14]，F の表象になる (Dretske, 1988, pp.95ff)．それでは，生物の内的状態 C は，いかにして F を表示する機能を得るのだろうか．

例えば，次のようなハトの学習を考えてみよう．ある装置は，信号が赤のとき（外部状態が F であるとき）かつそのときにのみ，バーを押すと餌が与えられるが，信号が青のとき（外部状態が G であるとき）にはバーを押しても餌は与えられない．ハトは，最初はやみくもに，あるいは偶然の仕方でバーをつ

つくだろうが，（以下では，ハトがバーをつつく際に生じる身体運動のタイプを M と呼ぼう）次第に，信号が赤のときにのみ，身体運動 M が生じるようになる．ドレツキによれば，ハトの内的状態 C はこのような学習を通じて，信号が赤であること (F) を表示する機能を獲得する（すなわち，F を表象するようになる）．

だが，学習を経て，C が F を表示する機能を獲得するとは，どのようなことだろうか．言い換えれば，ドレツキは，学習前から学習後においてハトにどのような変化が生じるのだと考えるのだろうか．

まず，学習前の時点において，ハトは生得的に，内部に，赤であること (F) を表示する表示子である，内的状態 C を持つ．学習前の時点では，ハトの内的状態 C と F（信号が赤であること）の間に，表示関係は成り立っているが，表象関係は成り立っていない．ドレツキによれば，赤の表示子を持つことで，ハトは知覚的に赤と別の色を見分けられるのだという [15]．ドレツキは，こうした F 表示子を持たない限り，生物は F の際にある行動をすることを学習できず，F を表示する機能は獲得されない，という．

次に，学習の途中段階において，ハトがやみくもに，またはランダムな仕方でバーをつついているような状況を考えてみよう．ハトはあるとき，偶然，信号が赤である際 (F) に，バーをつついた．このとき，外部状態が F であるときにハトに内的状態 C のあるトークン（例えば脳状態トークン）が生じ，それが偶然，バーをつつく際の身体運動 M のあるトークンを引き起こしたのだと考えられる．こうした学習の途中段階においては，C と M の間には，まだタイプ的な因果関係は成立していないと考えられる．つまり，M タイプのトークンは，C 以外のタイプのトークン（例えば外部状態が G である際に生じる内的状態 D）によって引き起こされることもたびたびあり，その意味で，C タイプのトークンによって M タイプのトークンが引き起こされるようなメカニズムは，まだ形成されていないと考えられる．また，このような学習の途中段階において，ハトの内的状態 C トークンと F のトークンの間に表示関係が成り立っているが，まだ表象関係は成り立っていない．つまり，生物の内的状態

C はまだ F を表象しておらず，C は F を表示する機能を獲得してはいないのである．

このような学習過程を経て，しだいにハトは，信号が赤である (F) ときにのみバーをつつくようになる．この意味で，ハトは，信号が赤である (F) ときにバーを押すと餌が得られることを学習する．このときハトの内部では，信号が赤であるとき（F タイプのとき）に形成される内的状態 C タイプによってバーをつつく際の身体運動 M タイプが引き起こされるというメカニズムが形成される [16]．ハトにそのようなメカニズムが獲得されたのは，F である際に生じる内的状態 C（つまり F と表示関係にある C）によって，身体運動 M（バーをつつく際の動き)が引き起こされたということが, ハトにとって有益な結果(餌)をもたらしたためであると考えられる．つまり，その有益な結果が，C が M を引き起こすことの強化子となり，こうした強化学習によって，先のようなメカニズムが形成されるのである．このメカニズムにおいて，内的状態 C は信号が赤であるときに形成されるべきものであり，「信号が赤であること (F)」を表すという役割を担わされている．したがって，C は「信号が赤であること」という外部の状態 F を表示する機能をもつと言える．

ちなみに，その他のメカニズムが選択されなかったのは，それらがハトにとって有益な結果をもたらさなかったからであると考えられる．例えば，外部状態が G である（信号が青である）際に生じる内的状態 D（つまり G と表示関係にある D）によって，M が引き起こされたとしても，餌は与えられず，そのようなメカニズムは選択されないと考えられる．

このように，ドレツキによれば，動物は，学習を通じて，外部状態が F である際に生じる内的状態 C が M を引き起こすようなメカニズムが獲得する．このようにして C は「F を表示する機能」を獲得する，すなわち F を表象するようになるのである [17]．

こうしたドレツキによる学習に基づく表象論は，進化に基づくミリカンの表象論同様，目的論的意味論に分類されるが，それは次のような理由のためであろう．すなわちドレツキが，生物の表象が，どのように形成され利用されるの

かということをふまえたうえで，行動を介して生物の生存に役立つものであると捉えているためである．その意味で，単に「表象するもの」と「表象されるもの」の関係のみを問題にしていたドレツキ初期の理論と比較し，表示機能・学習・行動の成功といった枠組みで表象を捉え直した『行動を説明する』における理論には，大きな変更が加えられたと考えられるだろう．

2.3 ドレツキによる誤表象の説明

それでは，こうしたドレツキの新たな表象論は，いかにして誤表象問題を解決に導いているといえるのだろうか．これまでにみたように，学習を経たハトの内的状態 C は F を表示する機能を持ち（内的状態 C は F を表象する），この機能は，うまく働くときとそうでないときがある．表示機能がうまく働くときには，C タイプのあるトークンは F タイプのあるトークン（信号が赤であること）を表示する[18]．なぜなら，このとき F は実際に存在しているからである．したがって，C は F を正しく表象しているといえる．1 節でみたミリカンの理論と同様に，ドレツキも，表象を行動の成功や失敗と結び付けて説明する．信号が赤のときに C が生じ，それによって身体運動 M（バーをつつく際の動き）が引き起こされたとすると，餌が与えられ，ハトの行動は成功であるといえるだろう．

一方，F を表示する機能がうまく働かないときには，C タイプのあるトークンは F を表示していない．それは，F が実際に存在しないにも関わらず（例えば，信号が赤ではなくて青であるにも関わらず），C のあるトークンが生じているときである．このとき F のトークンは存在していないので，C のトークンは F のトークンを表示しているとはいえず，誤表象している．このように，内的状態 C がうまく機能しないとき，つまり「本来表示すべきもの」を表示するということが成り立っていないとき，誤表象が生じることになる．誤表象は，行動の失敗と結び付けて説明される．信号が赤ではなく青のときに C が生じ，それによって身体運動 M（バーをつつく際の動き）が引き起こされたとすると，

ハトに餌は与えられず，ハトの行動は失敗に終わるだろう．

このようにドレツキは，機能がうまく働かない可能性と行動の失敗という観点から，誤表象を説明する．だが，こうした理論によって，ほんとうに誤表象を説明できているといえるだろうか．また，ドレツキの新たな表象論は，志向性を説明する理論として，ほんとうに問題のない立場なのであろうか．次節で検討しよう．

3. ドレツキ表象論の３つの問題点

前節でみたように，ドレツキによれば，「C が F を表象する」とは「C が F を表示する機能を持つ」ことにほかならず，生物の内的状態 C は行動における学習を経て，「F を表示する機能」を獲得することになる．そして，決して誤りえなかった単なる表示子（すなわち学習前の内的状態 C）は学習を経て，誤りうる表象になる．というのも，F を表示する機能を持つということは，それがうまく働かない可能性を含意するためである．誤表象は機能がうまく働かないときに生じるものであり，誤表象が生じるときには行動は失敗に終わると考えられる．

以上がドレツキによる表象及び誤表象への説明であるが，以下でみるように，こうしたドレツキ理論は，少なくとも，3 つの問題を抱えていると考えられる．後にみるように，これらはいずれも，ドレツキが厳しい「表示」概念を採用しているために問題であると考えられる．

3.1 誤りえない表示子がいかにして誤りうる表象になるのか

まずは，ドレツキによる「表示」がどのような概念であったのかを振り返ってみよう．前節でみたように，ドレツキによる「C が F を表示する」条件とは，C と F の間に，F タイプなしに C タイプが生じないという関係が成り立っているときである．こうした定義上，「誤表示」はありえず，ドレツキによれ

ば，学習前の動物はこうした誤りえない F 表示子（例：赤表示子）を持つ．例えばハトは，生得的に F（赤）を表示する内的状態，表示子 C をもち，この C とは赤の知覚的状態である．

しかし，ここで次のような疑問が生じる．動物は学習前であっても知覚的に誤ることがあるのではないだろうか．ドレツキの表示の定義によれば，「F なしに C は生じない」ことになり，これにしたがえば，学習前のハトにおいて，赤の知覚 (C) は，信号が赤であるとき (F) かつそのときにのみ生じることになる．しかし，学習前のハトにおいて，信号が赤でないとき（F でないとき）に赤の知覚 (C) が生じるようなこともあるのではないか．そうだとすると，ドレツキが，なぜハトが学習前から誤りえないような表示子を持つのだというのかは理解しがたい．

ドレツキは，ハトは赤の表示子を持つことで，赤と別の色を知覚的に区別できるのであり，それゆえ青ではなく赤の際にバーを押すということを学習するには赤の表示子が必要であり，こうした表示子なしに学習は成立しないのだという．もちろん，赤と別の色を知覚的に区別できない限り，赤の際にバーを押すことを学習することができないと考える点では，ドレツキは正しいだろう．ところが，表示子を持たない限りそのような学習は不可能であるというのは言い過ぎではないだろうか．学習は，表示関係なしにも生じると考えられる (Noordhof, 1996)．それなりに高い確率で F の際に生じるが，ときに F なしにも生じることがあるような知覚状態 C を考えてみよう．このとき，これらの間にドレツキのいう表示関係（F なしに C は生じないような関係）は成り立っていない．しかし，C が生じるときに高確率で F であり（信号が赤であり），その際に C によってバーをつつく際の身体運動 M が引き起こされ，餌を得られるのであれば，学習を経て F の際に C が生じ，C によって M が引き起こされるようなメカニズムがハトの内部に形成されることもあるのではないだろうか．

さらに重要なのは次の点である．学習前までは「F なしに C が生じなかった（C と F は表示関係にあった）」にもかかわらず，どうして学習後，「F なしに

C が生じる」ことが可能になるのだろうか（どうして誤表象が可能になるのだろうか）．言い換えれば，学習前は，「赤の知覚が生じるときには常に信号が赤であった」のに，どうして学習後は，信号が赤でないときにも赤の知覚が生じるようになるのだろうか．ドレツキによる説明では，学習前は単なる表示子であった C が，学習後は「F を表示する機能」を持ち，この機能がうまく働かないことがあるために誤表象が生じるということになる．しかしこうした説明は，F でないときには決して生じなかったはずの内的状態 C が，なぜ F でないときにも生じるようになるのかの説明にはなっていないだろう．

確かに，ドレツキが，誤表象を説明するためには，C と F の関係だけでなく，身体運動 M を考慮しなければならないと考えた点は正しい．これまでに見たように，ドレツキによれば，学習前の F と C の関係に，学習を経て M が結び付けられることで，正誤が生まれることになる．しかし，そうだとしても，学習前には F なしには生じなかった C が，学習後はいかにして F なしに C が生じるようになるのか（誤表象が生じるようになるのか）という点は，明らかでない．この点がうまく説明されない限り，ドレツキは依然として誤表象を説明できているとはいえないだろう．

3.2　表示の定義と表象の定義に関する問題点

次に，ドレツキ理論の第二の問題点を説明しよう．その前にまずは，ドレツキによる表象の定義をもう一度確認しておこう．ドレツキは「C が F を表象する」とは「C が F を表示する機能を持つ」ことであると定義する．こうした機能がうまく働くときには，C のあるトークンは（実際に存在している）F のあるトークンを表示することになる．一方，機能がうまく働かないときには，C のあるトークンは F が実際に存在していないにもかかわらず生じることになり，C は F を表示しないことになる．

しかし，表象のこうした定義には問題があると考えられる．というのも，「C が F を表示する機能」がうまく働くときさえも，ドレツキによる表示の定義上，

C トークンは F トークンを表示することが不可能になってしまうという問題が生じると考えられるからである．前述の通り，「C トークンが F トークンを表示する」条件とは，「F タイプなしに C タイプは生じない」という関係が成り立っていることである．だが，「C が F を表示する機能を持つ」場合，C の機能はうまく働かない可能性があることになり，C タイプのトークンは F タイプのトークンが実際に存在していないときにも存在しうることになる．そうだすると，このとき，「F タイプなしに C タイプは生じない」という関係が成り立っていないことになり，「C トークンが F トークンを表示する」ことが不可能になってしまうと考えられるのである．

それでは，なぜこのような問題が生じてしまうのであろうか．それは，ドレツキによる「C が F を表示する機能を持つ」という表象の定義には，「表示」と「機能」という両立しえない 2 つの概念が混在してしまっているためだと考えられる．言い換えれば，「F なしに C は生じない」という強い制約の表示の概念と，うまく働かない可能性を含む機能の概念は，折り合わず，それゆえ，矛盾が生じてしまうのだと考えられるのである．そうだとすると，ドレツキによる，表示と機能の概念の両方を取り入れた表象の定義には，問題があるといわざるをえないだろう．

3.3 生物にとって識別できない対象のケースにおける誤表象の説明

最後に，ドレツキ理論の第三の問題点を説明しよう．第三の問題点として，以下では，ドレツキが生物にとって識別できないケースにおいて誤表象をうまく説明できていないことを示す．そのために，まずは次のようなケースを考えてみよう．

> ロビンという鳥は，実験室のなかで，もともと ESA という，栄養になる人工的なエサを与えられ，食べていた (1).
> だが，あるときロビンの前に，ESA によく似た DOKU という毒である物質

が現れ，それを食べてしまった (2).

こうしたケースにおいて，(1) の時点までに ESA を食べる際に生じていたロビンの内的状態 C は，何を表示する機能を持つことになるのであろうか．「ESA を表示する機能」であろうか，それとも，「ESA と DOKU に共通する性質を持つ対象を表示する機能」であろうか．ドレツキにしたがえば，C が持つのは「ESA と DOKU に共通する性質を持つ対象を表示する機能」であろうと解釈できる．その根拠は，以下の通りである．

前述のとおり，ドレツキによれば，学習前の生物があるタイプの状態を表示する表示子を持たないなら（すなわち，あるタイプの状態を別のタイプの状態と知覚的に区別できないなら）内的状態はそのタイプの状態を表示しえず，学習を経てそのタイプの表示機能を獲得することができない．そうだとすると，ESA と DOKU を知覚的に区別できないなら，ロビンはもともと ESA 表示子を持っているとはいえない．それゆえ，ロビンの内的状態 C は「目の前に ESA があること」を表示することができず，「目の前に ESA があること」を表示する機能を獲得できないことになるのだと考えられる．そうだとすると，(1) 以前の段階で行われたと考えられる ESA を食べるようになるような学習の結果，ロビンに形成されたのは，「目の前に ESA と DOKU に共通する性質を持つ対象がある (F)」ときに生じる内的状態 C が，（食べる際の）身体運動 M を引き起こすメカニズムであり，内的状態 C は，このような F を表示する機能を獲得したのだと考えられる．

しかしながら，そうだとすると，次のような問題が生じることになる．すなわち，(2) のとき（ESA によく似た DOKU という有毒な物質を食べてしまったとき）でさえ，ロビンの内的状態 C は誤表象しているとはいえなくなってしまう．その根拠は，以下の通りである．もし C が「ESA と DOKU に共通する性質を持つ対象を表示する機能」を持つのだとすると，DOKU が目の前にあるときに C が生じてしまうとしても，機能は正常に働いていることになり，誤表象は生じていないことになる．さらに，C が有毒な DOKU を食べる際の

身体運動 M を引き起こすとしても，行動は学習した通りに生じていることになり，失敗であるとはいえなくなってしまうのである．

もちろんドレツキは，生物にとって知覚的に区別できるかどうかという点を重視しており，その意味で，エサとそうでないものを識別することは生物の識別能力の範囲を超えていると考えるのであろう．それゆえ，食べるべき対象でない対象を食べたとしても，それは誤りでも失敗でもないのだと考えるのであろう．しかし，目的論的機能の観点，すなわち生物の生存にとって有益なのはどのような機能かという観点からすれば，ロビンが ESA ではなく毒である DOKU を食べてしまうようなケースにおいて，本来であれば，ロビンは DOKU を ESA と間違えたのであり，誤って食べてしまったのだといわざるをえない．というのも，DOKU を食べるという行動は，毒を摂取することであり，結果としてロビンの生存に害を及ぼすような行動である．それゆえ，DOKU を食べてしまった際，その行動は失敗であるといわざるをえない．ドレツキ表象論は，こうした直観がうまく説明できず，それゆえ，誤表象のケースをうまく扱えないという問題があることになる．

4. なぜドレツキ表象論には問題が生じてしまうのか

前節でみたように，ドレツキは，表象（及び誤表象）をうまく説明できておらず，その根拠として，少なくとも 3 つの問題を抱えているということが明らかになった．要約すれば，その問題とは，以下のようなものであった．第一に，生物が学習前に，決して誤りえないような表示子を持つということがいかにして可能なのかが不明確であり，こうした「F なしに C が生じない」メカニズムが，いかにして学習後は「F なしに C が生じうる」ようになるのか（誤表象が生じるようになるのか）が不明確である．第二に，C が「F を表示する機能」を持つ場合において，この機能がうまくいくときでさえ「C が F を表示する」ことが成り立たないことになってしまう．第三に，生物にとって識別できない対象のケースにおいて，誤表象をうまく説明できない．

こうした3つの問題に共通するのは，いずれもドレツキが「表示」概念を採用しているために生じている問題であるということである．すなわち，いずれの問題も，ドレツキが，「F なしに C は生じない」という強すぎる条件に基づく「表示」概念を採用しているために生じている．こうした表示概念を採用している以上，ドレツキの理論は表象（及び誤表象）をうまく説明できないといわざるをえない．

しかし，少なくともドレツキ自身は，以前の著作で問題となっていた誤表象問題の解決を試みて，『行動を説明する』において新たな理論を構築したのだと考えられる．そして，既にみたように，ドレツキは少なからず，機能・行動・学習といった概念を導入することで，誤表象をうまく説明できると考えていた．だが，これまでにみたように，実際には，ドレツキによる新たな理論もまた，誤表象をうまく説明できているとはいいがたい．だが，そうだとすると，ドレツキはなぜ，機能・行動・学習といった概念を取り入れることで誤表象を説明できるという勘違いをしてしまったのだろうか．

それは，ドレツキが，表象するものと表象されるものの関係（内的状態 C と外部状態 F の関係）のみによって表象内容を説明するのではなく，C によって引き起こされる身体運動 M（と，C が M を引き起こすようになった学習過程）を含めた説明を行えば，行動の失敗という観点から，誤表象が説明できるのではないかと考えていたためであろう．つまり誤表象を，行動が失敗するようなケースにおいて生じるものとして位置付けることで，それを説明しようとしていたのがドレツキの戦略であったのだろう．ドレツキはこのような意味で，学習において内的状態 C が獲得した，「F を表示する機能」がうまく働かない際に，誤表象が生じると考えていたのである．

しかし，この点に関して，ドレツキは重要な勘違いをしていると考えられる．学習において生物が獲得したのは，内的状態 C と外部状態 F のつながりではなく，内的状態 C と身体運動 M のつながりのはずである．ドレツキの表示の定義にもあるように，C と F の間には，学習前からすでに法則的なつながり，すなわち「F でないときには C は生じない」という関係が成り立っているので

ある．そうだとすると，この関係は学習前から存在しており，学習を経てもCとFの関係には何の変化も生じていないことになる[19]．それでは，学習において生物に生じたのは，どのような変化だろうか．それは，（Fである際に生じる）Cが，身体運動Mを引き起こすようになったということである．そうだとすると，学習において生物のCが獲得したのは，「Fを表示する機能」というよりはむしろ，「Mを引き起こす機能」であるというべきであろう．（Fである際に生じる）Cは，学習過程において，まさにMを引き起こしてきたゆえに，生物にとって有益な結果をもたらしてきたのであり，それゆえにCはMを引き起こすようになった．それゆえ，CはMを引き起こす機能を獲得したのだというべきであろう．

先に述べたように，ドレツキは，行動の失敗という観点から，誤表象が説明できるのだと考えていた．しかし，行動が失敗するケースとは，せいぜい，Cの持つ「Mを引き起こす機能」がうまく働かないというケースにすぎず，「Fを表示する機能」がうまく働かないケースというわけではないと考えられる．そして，「Mを引き起こす機能」がうまく働かないというだけでは，誤表象の説明にはならない．というのも，誤表象を説明するためには，Fでないときに Cが生じるということがいかにして可能なのかを示さなければならないからである．しかし，これまでにみたように，ドレツキが，「Fでないときに Cは生じない」という表示関係を前提にした表象論を構築している以上，それは不可能なのである．

5. 結論と今後の課題——学習に基づく表象論に残された道

私たちの心の持つ志向性，すなわち心が別の何かを表象するという性質を説明するためには，いかにして誤表象が生じうるのかという問題が説明されなければならない．本論文では，学習に基づくドレツキ表象論を検討することで，志向性の自然化がいかにして可能なのか，特に，ドレツキが誤表象をきちんと説明できているのかという問題に焦点を当てて検討してきた．ドレツキ初期の

表象論は，表象するものと表象されるものの間の関係のみ，すなわち表象の生産の側面のみを考慮する理論であったために，誤表象を説明できないという問題を抱えていた．そこで，ドレツキは後に，『行動を説明する』における学習に基づく表象論において，機能・行動・学習という概念を取り入れ，表象を「F を表示する機能を持つ」ことであると定義することで，誤表象の説明を試みた．こうした新たなドレツキ表象論には，目的論的意味論の主流である，ミリカンによる進化に基づく表象論同様，表象の生産の側面のみならず，消費の側面が取り入れられてはいる．しかし，「表示」の概念を重視する限り，ドレツキ表象論は依然として生産重視であろうと考えられる．「表示」は「機能」の概念とうまく折り合わず，それゆえ，ドレツキの新たな表象論は結局のところ，誤表象を説明できていない．

それでは，こうしたドレツキによる学習に基づく新たな表象論は，志向性を説明する理論としては，一切の希望もないような理論だといわざるをえないのだろうか．そのように結論づけるのは早計であろう．少なくとも，「表示」の条件の制約を弱めることで，3 節で提示した問題をある程度克服し，誤表象問題を一歩でも解決に導ける可能性はあるだろう．例えば表示関係を，「F なしに C は生じない」というほど強いものではなく，ミリカンのようにある程度の相関関係で十分だと捉える方法がある (Millikan, 2004)．そうすれば少なくとも 3.1 項と 3.2 項で示した問題は回避できる可能性がある．だが，3.3 項で示した識別できない対象のケースにおける誤表象の問題は，ドレツキが表象 (C) の機能とは「F を表示する」ことだと考えている限り，生じるであろう．もちろん，4 節でみたように C の機能を「身体運動 M を引き起こす機能」として捉えるとともに，ミリカンのように表象の消費の側面をさらに重視するような表象論へと改訂するのであればそのような問題は生じないだろうと考えられる．とはいえ，表示の条件の制約を緩めることや，表象の機能を，身体運動を引き起こすことだと捉え直すことによって，ドレツキ理論に修正を加えると，その立場は，ミリカンの立場と大きく近づくことになる．その結果，修正されたドレツキ理論がミリカン理論とほんとうに異なるものであるのかという点が，新たな

問題として生じる．

もちろん，ドレツキ表象論の最大の特徴は，目的論的意味論の主流である進化に基づく説明ではなく，学習に基づく説明であるという点であろう．しかし，学習に基づく表象論が進化に基づく表象論とほんとうに異なるものであるかどうかという点には，検討の余地がある．というのも，ある表象を形成し利用するメカニズムが，ある種のレベルで選択され淘汰されるのか，それともある個体のレベルで選択され淘汰されるのかという違いに過ぎないとすれば，進化と学習に基づく表象論は，どちらも選択過程であるという意味では，差がないものであると考えられるからである．そうだとすると，学習に基づく表象論が進化に基づく表象論では説明できないような側面があることを明らかにしない限り，その必要性は主張できないだろう[20]．だが，こうした点に関しては，本論文の範囲を超えるので，今後の課題としたい．[21]

註

1　例えばこの後にみるように，信念をはじめとする志向性を持つ心的状態は，それ自身とは別の何かを表象する．とはいえ，全ての心的状態が志向性を持つわけではない．例えば，痛みなどの感覚的経験は，ふつう，志向性を持たないとされる．それゆえ，感覚的経験のように志向性を持たない心的状態は何も表象しないと考えられる．とはいえ，感覚的経験が志向性を持つという立場（例えば，Dretske, 1995）もあるのだが，感覚的経験が志向的であるかどうかという問題に関しては，本論文では扱わない．いずれにせよ本論文の関心は，志向性及び信念をはじめとする志向性を持つ心的状態にあり，志向性を持たない心的状態は扱わないものとする．

2　ドレツキが情報の概念を導入した本来の意図は，伝統的な認識論の問題を解決すること，特に知識を再定義することであったと思われる．実際ドレツキは，知識を，「情報によって引き起こされた信念である」として定義している (Dretske, 1981, p.86)．しかし，本論文では，認識論の問題は扱わない．

3　ここでの A と B の表象関係は，タイプ間の表象関係である．そのため，より正

確にいえば，「A が B を表象する」のは，A タイプの事象が B タイプの事象によって因果的に引き起こされ，A タイプの事象が B タイプの事象についての「情報を運んでいる」ときであり，そのときに限られる，ということになる．

4　もちろん心的状態そのものに目的論的機能を認めることはできない．というのも，例えば「目の前にヘビがいる」という個々の知覚は，ヘビがいなくなれば消失し，進化の過程で選択され存続するものではないからである．とはいえ，心的状態は，その形成機構と利用機構から「派生した」目的論的機能を持つと考えられる．

5　ドレツキの新たな表象論が誤表象問題を解決できているかどうかについて検討している数少ない論者 (Summerfield & Manfredi, 1998) さえも，学習が含まれるような事例をもとにした検討を行っておらず，議論が噛み合っていないと考えられる．

6　ドレツキは，更に後の著作 (1995) においてクオリアの自然化を目的とした理論も展開している．しかし，本論文の関心はあくまで『行動を説明する』における学習に基づく表象論にあり，後の理論は扱わないものとする．

7　学習に基づく表象論が進化に基づく表象論と本質的に異なるものであるのかという問題に関しては，検討が必要であろう（例えば，Dretske, 1990 と Millikan, 1990 を参照せよ）．というのも，ある表象を形成し利用するメカニズムが，ある種のレベルで選択されるのか，それともある個体のレベルで選択されるのかという違いに過ぎないと考えることもできるからである．こうした問題に関しては，註 20 で再び触れるが，本論文の目的は，あくまでドレツキの学習に基づく理論が誤表象をほんとうにうまく説明できているかどうかを検討することであり，この問題には深くは立ち入らない．

8　『行動を説明する』において，ドレツキは，「C が F を表示する」とは，前著での「C が F についての情報を運ぶ」を言い換えた表現であるという．しかしドレツキはこれらを全く同じ意味で使っていないのではないかという指摘もある (Millikan, 1990)．以下では少なくとも，『行動を説明する』においてドレツキがどのように表示の概念を定義しようとしていたのかを，後の論文 (Dretske, 1990) もふまえて，説明する．

9　本論文ではこのように，表示関係が成立するには，タイプレベルでの完全な相関が必要であるという解釈を採用する．確かに，『行動を説明する』では，表示関係が，

タイプレベルで成り立つ関係なのか，それとも単にトークンレベルで成り立つ関係（すなわちタイプレベルでの完全な相関を必要としないような関係）なのかは明確でない．もちろん本論文では前者の解釈をとるが，後者のように解釈することもできると思われるかもしれない．しかし，後の批判と応答 (Dretske, 1990) において，ドレツキは，表示関係が成り立つには，タイプレベルでの完全な相関が必要であることを明記している．また，もし仮に，表示関係が後者のようなものであると解釈するとしても，C が何を表示しているのかが不明確になってしまうという問題が生じる．というのも，C が生じている際に同時に世界の側に生じている事象は数多く存在するためである．以上のような理由から，本論文では，表示関係が成立する条件として，タイプレベルでの完全な相関が必要であるという解釈を採用する．

10　だが，ほんとうにそうだろうか．3 節で検討する．

11　だが，このとき C が F を表示することはほんとうに可能だろうか．3 節で検討する．

12　だが，3 節でみるように，こうしたドレツキによる説明には問題があると考えられる．

13　ドレツキにとって行動 (behavior) とは，あるシステムの内的状態 C が身体運動 M を生み出す過程である (Dretske, 1988, pp.2ff)．ここで注意を要するのは，行動とは，内的状態 C が身体運動 M を引き起こす過程 (C → M) であって結果 (M) ではないという点である．その意味で，行動は，行為 (action) とは区別される（例えば Davisdon, 1980）．

14　ドレツキが生物に与えているのは内在的表示機能と呼ばれるものであり，これは人工物が持つ表示機能からは区別される．ドレツキが生物に認めるのは，外からの解釈者を必要としない，それ自身が内部に持つ表示機能である．

15　なお，ドレツキによれば，ハトは学習前の時点では，赤と別の色を概念的には区別できないが，学習後，ハトは赤と別の色を概念的に区別できる．すなわち学習後，ハトは，餌に関するものとして，赤を他の色から区別できるのだとドレツキはいう．

16　もちろん，実際にそのような行動が生じるのは，ハトが空腹である場合，すなわち餌に対する欲求を持つような場合である．ドレツキは，内的状態 C を，人間で

いうところの信念に関する部分（F を表示する機能を持つ部分）と，欲求に関する部分に分けて更なる説明も行っている (pp.109ff)．しかし，ドレツキは，欲求を表象とはとらえておらず，それゆえ本論文では，こうした欲求に関する検討には深くは立ち入らない．欲求に関する議論は，例えば，Stampe, 1990 を参照のこと．

17　ところで，生物の内的状態 C のほかにも，身体運動 M の原因と呼びうる様々な候補がある．例えば C の原因である内的状態 B や，様々なホルモン及び神経学上の出来事としての内的状態 E などが想定できる．それでは，ここであえて C を原因として採用するのはなぜだろうか．それは，学習過程において，バーをつつく際の運動である M を引き起こし，ハトにとって有益な結果をもたらしてきたのは，まさに内的状態 C タイプのトークンが，信号が赤であること (F) を表示してきたゆえであると考えられるからである．

18　先の註でも述べたが，これが本当に可能であるかどうかに関しては，3 節で議論する．

19　もちろんドレツキは，学習を経て C は誤りうるようになり，その意味で C と F の関係は変化するのだというかもしれない．しかし，3 節でみたように，いかにしてその変化が生じるのかが説明されていない以上，その考えを受け入れることはできないだろう．

20　ドレツキは，生得的及び反射的な行動に用いられるような表象は進化による理論によって説明されるが，意図的な行動の理由になりうるような表象は学習による理論によって説明されるといい，その意味で自身の学習に基づく表象論の重要性を主張する．だが，進化と学習のどちらもある表象メカニズムの選択過程であるとすると，いかにして両者の理論が区別できるのかという点は明らかではない．少なくともドレツキの用いるハトの学習の例をみる限りでは，ハトが学習を経て意図的な行動をするようになったということや，意図的な行動の理由になりうるような表象を持つようになったということが示されていないと考えられる．ドレツキの説明によって示されたのはせいぜい，ある内的状態がある身体運動を引き起こすようなメカニズムが選択されたということにすぎないであろう．そうだとすると，進化に基づく表象論と学習による表象論が本質的に異なるものであるとは言い難い．ちなみ

にミリカンは，学習に基づく理論によって説明されるとドレツキが主張するような表象もまた，進化に基づく理論によって説明できると考え，その意味では両者の理論がそれぞれ別に必要であるとは考えていない．だが，こうした点に関しては，さらなる検討が必要であろう．

21　本稿は日本学術振興会特別研究員研究奨励金による研究成果の一部である．

文献

Davidson, D. 1980, *Essays on Actions and Events*, Oxford: Oxford University Press.［ドナルド•デイヴィドソン『行為と出来事』，服部裕幸，柴田正良訳，勁草書房，1990 年．］

Dretske, Fred. 1981, *Knowledge and the Flow of Information*, Cambridge, Mass.: MIT Press.

———. 1988, *Explaining Behavior: Reasons in a World of Causes*, Cambridge, Mass.: MIT Press.

———. 1990, “Reply to Reviewers,” *Philosophy and Phenomenological Research*, Vol.50, No.4, pp.819–839.

———. 1991, “Dretske’s Reply,” *Dretske and His Critics*, edited by Brian P. McLaughlin, Oxford: Basil Blackwell, pp.180–221.

———. 1995, *Naturalizing the Mind*, Cambridge, Mass.: MIT Press.

Fodor, Jerry A. 1984, “Semantics, Wisconsin style,” *Synthese*, 59, pp.1–20.

Millikan, Ruth G. 1989, “Biosemantics,” *Journal of Philosophy*, Vol.86, No.6, pp.281–297.

———. 1990, “Seismograph Readings for “Explaining Behavior”,” *Philosophy and Phenomenological Research*, Vol.50, No.4, pp.807–812.

———. 2004, *Varieties of Meaning*, Cambridge, Mass.: MIT Press.

Noordhof, Paul. 1993, *White Queen Psychology and Other Essays for Alice*, Cambridge, Mass.: MIT Press.

———. 1996, “Accidental Associations, Local Potency, and a Dilemma for Dretske,”

Mind & Language, Vol.11, No.2, pp.216–222.

Putnam, H. 1981, *Reason, Truth and History*, Cambridge: Cambridge University Press.

Stampe, Dennis W. 1990, "Desires as Reasons—Discussion Notes on Fred Dretske's "Explaining Behavior: Reasons in a World of Causes"," *Philosophy and Phenomenological Research*, Vol.50, No.4, pp.787–793.

Summerfield, Donna and Manfredi, Pat. 1998, "Indeterminancy in Recent Theories of Content," *Minds and Machines*, Vol.8, pp.181–202.