論 文 の 内 容 の 要 旨

論文題目　A reinforcement learning model of the basal ganglia system
— towards realization of higher brain functions —

（大脳基底核の強化学習モデル
— 高次脳機能の実現に向けて —）

氏 名　伊藤 秀昭

ヒトの脳と同程度以上に高性能な機械を実現するために、脳のアルゴリズムを研究し真似ることは有望である。特に推論やプランニング等といった、脳の様々な高次機能は是非とも実現されなければならない。このような様々な機能を包括的に実現することのできる可能性のある学習理論の枠組みとして強化学習が存在する。強化学習は報酬最大化の理論であり、一方高次脳機能は生命にとって報酬（食料など）を得るために存在する可能性があるからである。

脳においては大脳基底核系が強化学習に関与しているという仮説が提案されている。しかし実際には簡単な順序学習等で議論されており不明な部分が多い。そこで本論文ではまずサルの大脳基底核の神経活動と強化学習モデルとの比較を行った。タスクとして彦坂興秀教授の考案による 1DR/ADR タスクを用いた。これは memory-guided saccade task、すなわち予め視覚刺激（cue 刺激）によりサッケード眼球運動の目標位置を提示してその位置を記憶させておき、その後その位置へサッケードを行なわせるタスクであるが、ADR タスクにおいてはどの目標位置の方向（“cue 方向”）でも成功時に報酬が与えられるのに対し、1DR タスクではいずれか１つの方向（“報酬方向”）でしか報酬が与えられないというように報酬条件がコントロールされる。報酬条件以外は全く同じであるため、報酬条件の相違の影響を直接的に調べることができ、様々な現象が発見されている。本論文ではまずそれらの

現象が強化学習モデルによってどのように説明できるかを考えた。

始めに、黒質緻密部のドーパミン（DA）ニューロンの発火活動と、強化学習モデルの１つである TD（temporal difference）モデルとを比較した。DA ニューロンの発火活動についてはSchultzらによって報酬の予測誤差をコードしているという仮説が提案されており、TD モデルの TD 誤差に対応すると予想されていた（DA=TD 仮説）。ただしこれを疑問視する意見も存在した。そこで 1DR/ADR タスクにおいてこの予想が正しいか検討した。第一に、“within-block change”について検討した。1DR/ADR タスクはブロックデザイン、すなわち報酬条件一定の連続する約 60 トライアルを１ブロックとして、これを報酬条件を変えながら繰り返す形で行われたが、各ブロックの開始時に報酬方向は教示されず、サルは報酬の有無を数トライアル経験することでそれを知ることができる仕組みになっていた。ここで、DA ニューロンの活動は各ブロックの最初の数トライアルにおいて大きな変化を示した（このようなブロック内の変化を within-block change と呼ぶ）。すなわち、始めの数トライアルでは報酬が与えられるべきタイミングで反応が見られたが、その後そのような反応は消え、代わりに cue 刺激に対して反応するようになった。これはサルが報酬方向を理解したことによる変化であると考えることにより TD モデルでよく再現することができた。第二に、“post-reward order effect”について検討した。1DR タスクにおいて、cue 方向は擬似ランダムによって決定され、報酬を得なかったトライアルが何回連続したか（post-reward trial count）によって、その次のトライアルが報酬を得られるトライアルであるかどうかの確率が異なっていたが、DA ニューロンの cue 刺激に対する反応も、post-reward trial count によって異なっていた（これを post-reward order effect と呼ぶ）。この現象は DA=TD 仮説が定量的な意味で厳密に成り立っており、さらにサルが post-reward trial count に相当する情報を保持しているとすると、よく再現できることが分かった。また、サルが報酬の予測誤差を近似的に計算していると仮定すると、さらに細かい部分まで再現できることも分かった。これらの結果により DA=TD 仮説の妥当性が支持された。

次に、基底核の尾状核（caudate; CD）ニューロンの発火活動と、強化学習モデルの１つである actor/critic モデルとの比較を行った。CD ニューロンについては、Houk らにより、actor/critic モデルによって説明できるという仮説（CD=actor/critic 仮説）が提案されていた。ただし実際の発火活動との比較はあまりなされておらず、また不必要に複雑なモデルが用いられたりしていた。そこで 1DR/ADR タスクにおいてできるだけシンプルな actor/critic モデルを用い CD ニューロンの発火活動がどの程度説明できるかを検討した。第一に、cue 提示前の発火活動について検討した。多くの CD ニューロンにおいてこの活動には報酬方向に対する選択性が見られ、その選択性がブロック内で徐々に顕著になる現象が見られたが、これは報酬方向を入力として与えられた critic モデルのニューロンにより再現できた。第二に、cue 提示後の発火活動について検討した。多くの場合報酬の有無に対する選択性が見られ、加えて一部で cue 方向に対する選択性も見られることが報告されて

いたが、それぞれ適当な入力を与えられた critic モデルのニューロンにより再現された。第三に、saccade 時の発火活動について検討した。ここでは多く場合報酬の有無および cue 方向に対する選択性が見られ、報酬の有無に対する選択性がブロック内で徐々に顕著になる現象が見られたが、これは cue 方向を入力された actor モデルのニューロンによって再現できた。第四に、以上の発火活動において post-reward order effect が見られたが、これは post-reward trial count を入力された critic モデルのニューロンにより再現された。これらによって代表的な CD ニューロンの発火活動の特徴が説明され、CD=actor/critic 仮説の妥当性が支持された。最後に、cue 提示後に報酬方向の情報を入力された critic モデルのニューロンは実際の CD ニューロンにおいて観察されなかった活動パターンを示した。これは報酬方向の情報が cue 提示後に CD から消失していることを強調する結果であり、基底核での報酬関連処理の特徴を示唆している。

このような特徴についてより詳しく調べるために解析を進めた。1DR タスクにおいては報酬方向と cue 方向とが存在し、cue 方向が報酬方向と異なる場合には報酬は与えられない。しかしその場合でも正しくサッケードを行わなければ同じ cue 方向のトライアルが繰り返されるため、いずれ報酬を得るためには正しく cue 方向へサッケードを行うことが必要である。よってこの場合、報酬方向へのサッケードをゴール、cue 方向へのサッケードをサブゴールと考えることができる。このゴールとサブゴールがどのように処理されているかは興味深い問題である。

そこでまずエラーサッケードの性質を調べたところ、報酬方向と cue 方向のいずれかの方向へのエラーが多く、cue 提示後時間の経過と共に cue 方向へのエラーの割合が増していた。これにより、報酬方向と cue 方向の両者が行動に影響を与えること、および cue 提示の前後で前者から後者へと行動のバイアスが変化することが示唆された。次に、行動と CD ニューロンの発火活動との関係を調べた。眼球運動のパラメータと CD ニューロンの発火頻度との間には実際の眼球運動の実行時よりもかなり前から有意な相関が見られた。これにより CD ニューロンが眼球運動の制御に関与していること、およびその関与は実行時の直接的なものだけではないことが示唆された。さらに、相互情報量解析により CD ニューロンの発火活動に含まれる情報とその時間変化を調べると、CD ニューロンの発火活動に含まれる情報は cue 提示の前後で報酬方向から cue 方向へと変化していた。これらの結果を総合し、CD では行動を制御するために報酬方向（ゴール）と cue 方向（サブゴール）が処理されており、それらが cue 提示の前後で前者から後者へと変化することが示唆された。この変化は cue 提示という外的イベントにより引き起こされたものと考えられるが、もし内的に（外的イベントに依存せず）同様のことが起こるとすれば、CD においてゴール指向性推論が行われる可能性を示唆している。

最後に、これらの結果を踏まえ、actor/critic モデルに基づいたゴール指向性推論モデルを提案した。まず、サブゴールの設定を actor の行動と考え、actor/critic モデルに組み入れた。しかしこれだけでは学習速度が遅かったので、“悪いサブゴールの拒否”と“二

重学習”という二つの改良を行った。前者はサブゴールを設定した結果報酬の予測値が減少した場合（すなわち TD 誤差が負の場合）にその設定を取り消すようにしたものであり、後者はサブゴールのサブゴールまで考慮して学習を並列に行わせるものである。これらにより、actor/critic モデルによってゴール指向性推論が実現され、学習性能も向上しうることが分かった。また、このような学習により階層的な行動制御が自動的に獲得され、さらに学習を重ねるとそれが消滅するという現象も示された。

以上により、大脳基底核系における報酬関連処理の特徴およびゴール指向性推論という高次脳機能への関与の可能性が示された。より詳細なモデルの検討、および大脳新皮質を始めとする大脳基底核系以外の脳部位との関係を考慮することがこれからの重要な課題である。