2A1-4

計算神経科学における強化学習
「神経修飾物質系のメタ学習仮説」

Reinforcement Learning and Computational Neuroscience
- Possible Functions of Neuromodulators in Metalearning -

銅谷賢治[*1,*2]
Kenji Doya

[*1] ATR 人間情報科学研究所
ATR Human Information Science Laboratories

[*2] 科学技術振興事業団 CREST
CREST, Japan Science and Technology Corporation

The framework of reinforcement learning captures an essential function of the nervous system: to realize behaviors for acquisition of reward. Thus the architectures and algorithms of reinforcement learning can provide important clues as to the organization and functions of the nervous system. Here I report three such examples: 1) a model of the basal ganglia as the circuit for reinforcement learning; 2) understanding of the specialization and collaboration of the cerebellum, the basal ganglia, and the cerebral cortex; 3) working hypotheses about the roles neuromodulators in regulating the metaparameters of reinforcement learning. The concept of reinforcement learning can provide a common ground for interdisciplinary studies.

エサなどの報酬を獲得するための行動を探索的に学習するという強化学習の問題設定は，動物や人間の行動学習の最も基本的な側面を捉えている．そもそも「強化」あるいは “reinforcement” という言葉は，心理学や生物学の世界から借りて来たものであるが，強化学習の工学的な研究の中で開発されたアルゴリズムやアーキテクチャは，人間や動物の行動学習の脳内メカニズムを理解する上で，有用な手がかりを与えてくれる．

1. 大脳基底核の強化学習モデル

大脳皮質と脳幹の間に位置する大脳基底核は，その病変により起こるパーキンソン病やハンチントン舞踏病などの症状から，運動制御に何らかの形かかわることが知られていたが，その正常時の機能は長く神経科学の謎であった．

その解明の糸口を与えてくれたのが，中脳から大脳基底核や前頭葉に線維を伸ばし，ドーパミンという物質を放出するニューロンの活動記録である[1]．LED が点灯したらレバーを押すという行動をサルに学習させた時，その初期にはドーパミンニューロンは報酬であるジュースに対して応答する．しかし行動学習が確立した後には，ドーパミンニューロンは LED の点灯に対して応答するようになり，ジュースに対しては応答しなくなる．これは，強化学習の理論で報酬の予測からの増減を表す「TD 誤差」

$\delta_t = r_t + V(s_t) - \gamma V(s_{t-1})$

の振る舞いと非常に良く似ている．TD 誤差は報酬予測と行動学習の両者の学習信号であり，この発見を契機に，大脳基底核とドーパミンニューロンの機能を，強化学習の枠組みで説明するモデルが提案されている（図１）[1,2]．ドーパミン系の賦活が薬物依存など行動の強化につながるという事実は以前から知られていたが，大脳基底核のシナプスの可塑性がドーパミンによって制御されることが実験的にも確かめられ，報酬の予測と，それに基づく行動選択が，大脳基底核の基本的な機能として理解されるようになった．

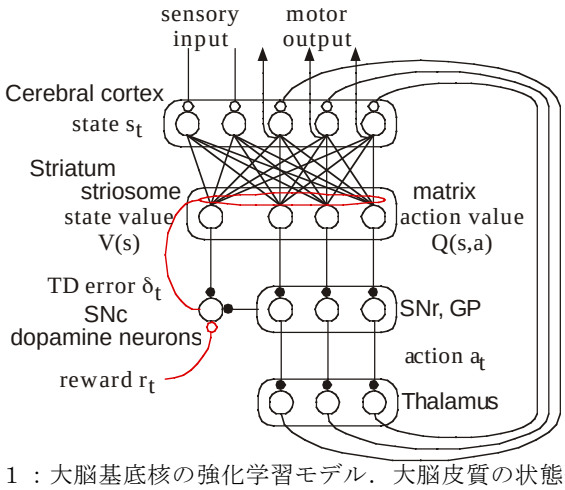

図１：大脳基底核の強化学習モデル．大脳皮質の状態表現 s_t をもとに線条体(striatum)で状態価値関数 V(s)と行動価値関数 Q(s,a)が計算され，黒質網様部(SNr)と淡蒼球(GP)を経て視床(Thalamus)や脳幹の運動中枢への回路で行動 a_t が選択される．黒質緻密部(SNc)は TD 誤差δ_t を線条体にフィードバックし，ドーパミン依存性のシナプス可塑性により価値関数が学習される．

2. 小脳，大脳基底核，大脳皮質の機能分化

PET や fMRI などによる人間の脳活動計測により，以前は主に運動に関わると考えられて来た小脳や大脳基底核が，言語処理や暗算，イメージ操作や表情認知など，運動以外の機能にも関わることが明白になった[3]．また，神経連絡を調べる技術の進歩により，小脳も大脳基底核も，高次認知機能の座であるとされる大脳皮質の前頭前野に出力を送ることが明らかにされた．そこで，運動制御だけではない小脳と大脳基底核の役割に関して，新たな思考の枠組みが必要とされている．大脳基底核を強化学習を実現するため

連絡先：銅谷賢治，ATR 人間情報科学研究所，619-0288 京都府相楽郡精華町光台 2-2-2，Fax: 0774-95-1259，E-mail: doya@atr.co.jp http://www.atr.co.jp/his/~doya

の回路するモデルは，さらに脳全体の機能分化の原理に関しても重要な示唆を与えてくれる．

これまでの研究で，小脳の学習は誤差信号をもとにした「教師あり学習」の枠組みで説明されている．また大脳皮質のニューロンの応答特性は，入力の統計的性質をもとにした「教師なし学習」の枠組みで説明されている．これらと合わせると，小脳，大脳基底核，大脳皮質はそれぞれ，教師あり学習，強化学習，教師なし学習という，異なる種類の学習のアルゴリズムに応じて専門化した神経回路であるという枠組みが浮かび上がる（図２）[4].

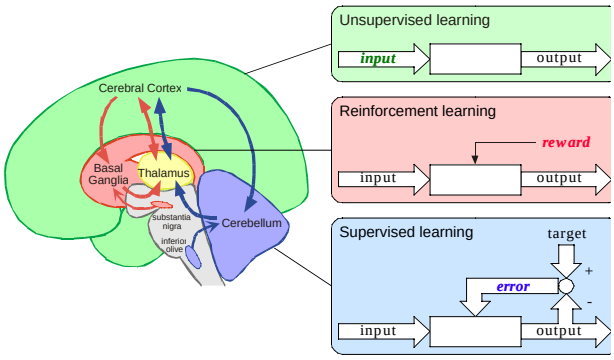

図２：小脳(cerebellum)，大脳基底核(basal ganglia)，大脳皮質(cerebral cortex)はそれぞれ，教師あり学習(supervised learnning)，強化学習(reinforcement learning)，教師なし学習(unsupervised learning)に特化した回路構造とシナプス可塑性のメカニズムを持つ．

例えば部分観測の強化学習では，価値関数による基本的な強化学習の要素に加え，隠れ状態の推定値(belief state)を更新する方式が用いられる．このような処理は，小脳に教師あり学習で獲得された環境の予測モデルと，大脳基底核に強化学習で獲得された価値関数を，大脳皮質に教師なし学習で獲得された隠れ状態表現を介してつなぎ合わせることで実現可能なはずである．このような脳の大局的な機能分化と機能統合に関する仮説は，fMRI などで得られる「脳のどこが活動している」というデータを「そこでどういう処理が行われている」という理解につなげる上で有用な手がかりとなり得る．

3. 神経修飾物質系のメタ学習モデル

強化学習をロボット制御などに適用してみて[5,6]痛感することは，ロボットが学習する以上にさせるために，実験者自身が学習のさせ方を学習させられているということである．強化学習アルゴリズムでは，将来の報酬予測の割引率(γ)，探索のランダムさを決める逆温度(β)，学習の速度係数(α)などの「メタパラメタ」の適切な設定は，学習課題や環境条件に依存するため，多くの場合，実験者の試行錯誤によるチューニングを必要とする．これが，研究室レベルでは成功を収めている学習ロボット達が，なかなか世の中に出て行けない大きな理由である．

一方われわれの脳は，その学習のメタパラメタを誰かにチューニングしてもらわなくても，未知の環境のもとで様々な行動を自律的に学習することが可能である．つまり脳には，その学習のメタパラメタを自己調節する「メタ学習」の機構が備わっていると考えられる．

脳におけるメタ学習の担い手として，脳幹から大脳皮質や基底核，小脳に広く投射し，持続的な作用を及ぼす「神経修飾物質系」が考えられる（図３）．近年，分子生物学

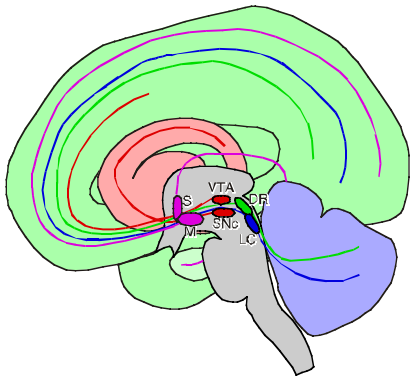

図３：代表的な神経修飾物質系：黒質緻密部(SNc)と腹側被蓋野(VTA)のドーパミン系，背側逢線核(DR)のセロトニン系，青斑核(LC)のノルアドレナリン系，中隔核(S)とマイネルト核(M)のアセチルコリン系．

的技術の発達により，様々な神経修飾物質とその受容体の脳内での分布や細胞レベルでの作用，またその阻害薬や遺伝子ノックアウトによる行動への影響に関するデータは膨大に得られている．それらの実験的知見と理論モデルをベースに，

1) ドーパミンは報酬予測からの増減（TD 誤差δ），
2) セロトニンは報酬予測の時間スケール（割引率γ），
3) ノルアドレナリンは行動のランダムさ（逆温度β），
4) アセチルコリンは記憶の更新の（学習速度α），

をそれぞれ表現し制御しているという仮説が考えられる[7].

この仮説の検証に向けて，CREST「脳を創る」の研究課題として，学習理論研究，ラットやサルでの生理実験，人間の脳活動計測，ロボット実験を組み合わせた共同研究が進んでいる．

4. おわりに

強化学習の問題設定は，より良い性能を求める工学にも，また生物の適応や人間の行動の原理に迫る研究にも共通の枠組みであり，異なる学問分野を貫く新たな研究の基盤を与えるものとなり得る．

参考文献

[1] Schultz, W., Dayan, P., and Montague, P.R.: A neural substrate of prediction and reward. *Science*, 275, 1593-1599 (1997).
[2] Houk, J.C., Adams, J.L., and Barto, A.G.: A model of how the basal ganglia generate and use neural signals that predict reinforcement. In J.C. Houk, et al. Eds: *Models of Information Processing in the Basal Ganglia*, pp. 249-270. MIT Press (1995).
[3] Doya, K.: Complementary roles of basal ganglia and cerebellum in learning and motor control. *Current Opinion in Neurobiology*, 10, 732-739 (2000).
[4] Doya, K.: What are the computations of the cerebellum, the basal ganglia, and the cerebral cortex. *Neural Networks*, 12, 961-974 (1999).
[5] Doya, K.: Reinforcement learning in continuous time and space. *Neural Computation*, 12, 219-245 (2000).
[6] Doya, K., Kimura, H., and Kawato, M.: Neural mechanisms of learning and control. *IEEE Control Systems Magazine*, 21(4), 42-54 (2001).
[7] Doya, K.: Metalearning and neuromodulation: *Neural Networks*, 15(4), (2002)