題目：報酬構造学習―ドーパミン神経細胞をめぐる新仮説―

Title: Reward structural learning by dopamine activity

中原裕之
Hiroyuki Nakahara

理化学研究所 脳科学総合研究センター 理論統合脳科学研究チーム〔〒351-0198 埼玉県和光市広沢 2-1〕
Laboratory for Integrated Theoretical Neuroscience, Riken Brain Research Institute, 2-1 Hirosawa, Wako City, Saitama 351-0198, Japan

はじめに

私たちの日常生活は、朝食に何を食べるか、職場で上司に何を報告するかなど、意思決定の連続である。意思決定は未来に影響を与える。逆に言えば、未来予測が意思決定の際には働く。よって、未来予測の学習は意思決定において本質的な役割を果たすことになる。経験から予測を修正、つまり学習することが、適応的な意思決定を可能にするのである。予測の中でも、報酬の予測はヒトや動物にとって原初的な部類に属する。報酬予測は、動機づけ行動の土台ともなる[1-4]。そのため、報酬予測の学習とその意思決定（価値に基づく意思決定：value-based decision making；以下、「価値意思決定」）における脳機能の解明は、神経科学の重要なテーマとなっている。近年、この価値意思決定に関する脳研究が、報酬予測の脳計算理論（強化学習：reinforcement learning theory）に支えられ発展してきた。そこでは、いわゆるドーパミン報酬予測誤差仮説が大きな役割を演じている[5]。本稿では、私たちが先ごろ新たに提唱した「ドーパミン報酬構造学習仮説」を紹介する。

背景：強化学習理論の枠組みとその前提

価値意思決定の脳機能解明の発展には、実験と理論研究の相互交流やその融合研究が大きく寄与してきた。その土台となったのが強化学習理論（より正確には、特に「時間誤差学習：temporal difference learning」）である[6]。この理論が本来的にもつ強みは、報酬予測の学習についての明確で、かつ数学的に定義された記述である。一方で、神経科学でこの理論が影響力をもつようになったのは、その本来的な強みよりも、むしろ、その理論適用に透明性があること、つまり、理論を実際の動物の行動や神経活動に適用するときの一連の明確な前提的仮定をもっていることによる。この仮定が強化学習理論で用いられる要素と脳メカニズムのマッピングを可能とし、この分野に大きな進展をもたらした。

この前提的仮定を理解するには、この理論を神経科学的実体とは切り離して考えることが大切である。それには、この理論の数学的構成を理解するのが一番だが、ここではそれよりも直観的理解を優先しよう。まずは、入力を受けて出力を出す「抽象的な存在」を考える。この「存在」は何でもいい。たとえば、インターネット上の一つのハブでもいい。「何でもいい」ことを強調するために、これを「存在」と呼ぶ。さて、入力に応じて出力が出されると、それがこの「存在」の外界に影響を与え、それによって次の入力が生まれる。その新しい入力に対して、この「存在」は次の出力を行う（以下、続く）。次の入力に移行するとき、この「存在」はある「数値」を受け取る。この「数値」は大きければ大きいほど「存在」にとって好ましい。「存在」は、この入力―出力のペアを繰り返し経験していく中で得られる「数値」の最大化に関心をもっている。与えられた入力に対して、どの出力を出したらどれだけの数値が得られるか（つまり数値予測）、そしてどの出力が一番よいか（つまり出力の選択）、この 2 つを学習するのが「存在」にとっては重要である。これが強化学習の構図の基本概念である。

この「存在」について具体的に考えてみよう。もっとも簡単なのは、「存在」をある環境（または実験課題中）の個体（ヒトまたは生物）と見なすことである。「入力」はその環境の状態である（以下、「状態」）。「出力」は個体からの環境への働きかけである。この働きかけを以下「行動」とよび、出力の選択を「行動選択」とよぶ。「数値」は「報酬」のように個体にとって重要なものである。この事例は確かに有用であるので、以下の記述でもこれらの用語を使う。

しかし、ここで、「存在」が“完結した個体”であると考えてしまうと、それは誤解である。強化学習理論においては、「存在」は入出力に対して報酬予測と行動選択をする実体であればよい。特に、強化学習が脳機能の一部にあるとみなしたときには、むしろ、その「存在」にとっての環境は個体にとっての外部環境とは一致しない。言い換えれば、強化学習を行う脳部位にとっての入力は、外部環境からの入力ではなく、むしろ、その部位に投射を行う神経細胞群の活動によって表される入力と考えるべきである。この点をより掘り下げて次に論じていきたい。

誤差情報と入力表現が強化学習の土台

上に述べた強化学習による報酬予測の能力は、報酬予測誤差によって表される情報の精度と、予測学習に影響する入力表現の豊かさによって決まってくる。話を簡潔にするために、時間誤差学習ではなく、その簡易版（レスコーラ＝ワグナーの学習則に相当する）を用いて、以下でこの重要なポイントを概観する。

強化学習で報酬の予測を行い学習するには、与えられた入力（状態）に対して、経験から予測を可変できる必要がある。報酬予測（価値関数とも呼ばれる）を$V(t)$として

$$V(t)=\sum_{i=1}^{n} w_i s_i$$

と表される場合を考えよう。右辺の $s_1,\ldots,s_n$ は、報酬予測のための入力を表している。そして、w_i は可変な重みづけである。学習信号となる報酬予測誤差 $\delta(t)$ は、「実際の報酬－予測した報酬」であり、たとえば、実際の報酬を $r(t)$ で表すと、

$$\delta(t)=r(t)-V(t)$$

となる。この報酬予測誤差が学習信号になる。報酬予測の学習は、報酬予測の誤差を修正する方向に変えていくことで実現される。つまり、

$$V(t)\rightarrow V(t)=V(t)+\Delta V(t)=V(t)+\alpha\delta(t)$$

となる。なお、各入力 s_i に対する重みについては、

$$w_i \rightarrow w_i = w_i+\alpha s_i(t)\delta(t)$$

となる。この学習を繰り返すことで、与えられた入力に対する報酬予測を正確にできるようになる。学習は誤差が（平均して）ゼロになるときに終了となる。ここで、誤差が学習を規定するのは自明なことだろう。なぜなら、学習は誤差がゼロになるときに終了するからである。ちなみに、行動選択の学習にも同様の考え方を適用することができる。

さて、報酬予測の学習を規定するもう一つの要素は、予測のための入力表現である。この点は重要であるにもかかわらず、しばしば見落とされがちである。簡単な例で考えてみよう。たとえば、外部からの入力が「赤」と「青」だったとして、それぞれの場合で報酬予測を学習したいとしよう。その赤と青がお互いに区別できる入力表現、たとえば、$(s_1,s_2)=(1,0)$ と $(s_1,s_2)=(0,1)$ で表されていたら、それぞれの色に対して予測の学習が可能である。しかし、この両方とも $(s_1,s_2)=(1,0)$ という同じ表現だったとしたら、赤と青の報酬予測を区別して学習することは不可能になる。わかりやすく言えば、入力情報が外部情報を十分に豊かに表現できているかどうかが、報酬予測の学習能力に影響するのである。

このように、報酬予測に関わる情報の入力表現が、報酬予測の精度を規定する。そうして生まれた報酬予測が、学習信号としての誤差によって表現される情報にも影響する。言い換えれば、学習される報酬予測自体が、その学習信号と予測の入力表現で決まってくるということになる。入力表現は、報酬予測においてまるで黒子のような重要な役割を担っているのである。

ドーパミン報酬予測誤差仮説の有用性と限界

強化学習理論が神経科学で注目を浴びたきっかけは、ドーパミン報酬予測誤差仮説の提唱であった[5，7-9]。ここで、その骨子を簡単に見ておこう。報酬予測誤差仮説は、ドーパミン神経細胞活動が、「時間誤差学習のもとでの学習信号である報酬予測誤差を表す、そして、その学習信号をもとにして報酬予測が学習される」という仮説である。この仮説の前半部を正確

に理解するには、時間誤差学習を理解した上でドーパミン神経細胞活動への適用を考えるべきだが、紙面の都合上それは省略する（引用文献[10]などを参照願いたい）。仮説の後半部「その学習信号をもとにして報酬予測が学習される」は、前節の考え方を援用すれば理解できる。脳の別の部位で「報酬予測」が生成され、ドーパミン神経細胞が「誤差＝実際の報酬－予測した報酬」という学習信号を表す。この学習信号がその予測する脳部位に投射され、そこで「重み」を変化させることで、報酬予測がより「実際の報酬」に近くなるように学習すると考えられたのである。

この仮説は、その提唱後、価値意思決定の脳機能解明に著しい発展をもたらした。その重要性は強調しても強調しすぎることはない。また、仮説はさしおいても、ドーパミン神経細胞が報酬予測の誤差を表すという実験的知見も、その後、何度もさまざまな実験で証明された。つまり、ドーパミン神経細胞活動を引き起こす主要因が、報酬予測誤差（または誤差らしきもの。これについては後の議論を参照）であるという生物学的知見が積み上げられていることを忘れてはならない。しかし、この仮説の有用性にはじつは限界があることも指摘したい。なぜなら、その限界を見極めることこそが、今後の価値意思決定の脳機能解明の発展につながるからである。

報酬予測誤差仮説で仮定する入力表現は、直近の外界からの感覚入力そのものである（図 1A）。脳内で生成される入力表現はそれには含まれていない。仮説提唱後の進展を振り返れば、脳内入力表現を無視することで、仮説の検証、つまり外界事象に基づく入力表現と脳内の報酬予測との対応づけがシンプルになるというメリットがあったのは確かである。一方で、外界事象を入力表現とすることは、その報酬予測の能力が外界事象によって制限されていること、また報酬予測の生成が、外界事象に対して受動的にしか行われないことを意味する。学習に使われる「重み」の変化以外には、脳内で生成される情報や入力表現には何も役割が想定されていないのである。

すなわち、報酬予測誤差仮説で想定されているドーパミン神経細胞の報酬予測誤差とは、直近の外界事象を入力とする特定の報酬予測を前提としている。では、ドーパミン神経細胞の報酬予測誤差は、本当に直近の外界事象にだけ基づく誤差なのだろうか。もしそうでないとしたら、それは何を意味するのだろうか。また、入力表現が直近の外界事象よりも豊かであれば報酬予測はどのように変わるのか。これらの疑問について次節で論じたい。

予測に有用な報酬構造：ドーパミン神経細胞の活動は報酬構造に影響されるか?

私たちは、実験と計算論的モデリングを融合した研究を行うことで、ドーパミン神経細胞活動が、報酬予測誤差仮説で想定されている報酬予測よりも優れた報酬予測をもとに、報酬予測誤差を表現しうることを発見した[11]（ここではその詳細は省くが、中原（2013）や理化学研究所（2004）による簡単な紹介がある[12, 13]）。要点だけ述べれば、私たちはある課題を用いた次のような実験を行った。その課題では、何回も試行を繰り返すな

かで、平均では 4 回に 1 回報酬がもらえるようになっていた。つまり、各試行において、その試行での入力だけを考慮した場合、報酬をもらえる確率は25%に設定されていた。したがって、報酬予測誤差仮説の報酬予測も当たる確率は 25%となる。一方、この課題では、その直前の試行で無報酬の回が続くほど報酬確率が高くなるように工夫がしてあり、もしこれらの試行での報酬の有無を考慮に入れた報酬予測を行えば、より精度の高い予測ができるようになっていた。つまり、その精度の高い予測をドーパミン神経細胞が利用していたら、その誤差信号はより精度の高いものになるはずだ。そして、私たちは確かにドーパミン神経細胞がその精度の高い予測を使った誤差を表現していることを見出したのである。つまり、ドーパミン神経細胞の誤差信号は、報酬予測誤差仮説の報酬予測よりも優れた予測を利用しうるとの結論を得た。この見解は、他の研究でもほぼ支持されている[14-19]。

では、ドーパミン神経細胞が表す誤差信号が、誤差仮説の想定よりも優れた信号であるということは何を意味するのだろうか。それは、その誤差を利用して行われる報酬予測が、より優れた報酬予測になりうることを意味する。一般的に言えば、その報酬予測は、直近の外界事象を超え、その外界環境における報酬の情報をより広くとらえた、つまり報酬環境構造を反映したものになりうることを示唆する。

実際、そういった報酬環境構造を反映することで、報酬予測が良くなる例はいくらでもある。たとえば、餌を探す時に、同じ木から果物を獲得するような場合には、報酬（果物）が得られたからと言って報酬予測を上げ続けるのは得策ではない。むしろ報酬が得られていく中で、いつかは得られる報酬が減っていく（つまり報酬予測を下げていく）ほうが妥当である。そして、どこかで見切りをつけて別の木に移るほうが得策である[20, 21]。一般に、直近の外界事象だけに囚われずに、それ以前に起きた事柄をうまく反映する、その環境の報酬構造をうまく利用する、といった報酬予測のほうがより精度が高くなる。アイスホッケーでの有名な逸話だが、パックが今あるところにいくのではなく、これからパックが行きそうなところに動く方が、よりゴールを生む可能性が高くなる。直近の外界事象からの報酬予測だけに囚われずに、これから起こりうる未来を予測してそれを報酬予測に取り込むほうが、予測が良くなる事例は多いのである。

ドーパミン報酬構造学習仮説

私たちは最近「ドーパミン報酬構造学習仮説」を提唱した（図 1B）[18]。これは、報酬の構造とその予測の学習は本来不可分であるという考え方を出発点としている。ドーパミン神経細胞活動は、誤差仮説で想定される報酬予測誤差信号を超えて、環境構造を反映する学習信号であり、それは予測の学習だけでなく環境構造を入力表現に反映するのにも適していると私たちは考える。すなわち、ドーパミン神経細胞活動は誤差仮説で想定する報酬予測のための「重み」の学習だけでなく、予測の入力表現の学習にも利用

される、つまり、予測学習と表現学習の両方に影響を与えるとするのがこの仮説である。

この構造学習仮説を支持すると思われるいくつかの根拠を挙げてみよう。第一に、たとえば知覚学習の分野では、そもそもドーパミン神経細胞が表現学習にも重要な役割を果たすことは多くの研究によって支持を受けている[22-24]。第二に、ドーパミン神経細胞活動の意外な多様性と、その多様性が表現学習に適している点が挙げられる。誤差仮説ではドーパミン神経細胞活動が、誤差信号として比較的一様であることが想定されていた。誤差仮説に触発されて、ドーパミン神経細胞活動のより詳細な知見が積み上げられてきたが、それらが指し示すのは、より多様な情報がその活動に表されていることである。上述の環境構造を反映した誤差信号の他に、たとえば、「不確実性」[25]や「事前情報」[26]、「運動開始」、「アラート信号」そして「サリエンス信号」などの情報がある[15, 27, 28]。また、予測と行動選択の観点からは、探索に関わる情報や、あるいは、行動選択や課題構造に関わる情報が、報酬予測誤差信号に付け加わることがあることも示されている[29-31]。特記すべきは、これらの活動はすべて、原理的に表現学習を助けることができる点である。第三に、ドーパミン神経細胞の線条体などの基底核関連回路や前頭葉などの大脳皮質への広い投射を考えると、その学習信号としての役割が、そもそも予測だけに(つまり入力表現が与えられた後の予測のための「重み」の学習だけに)関わっているとするのは考えづらい。むしろ、予測と表現の両方の学習に関わると考える方が自然である。第四に、ドーパミン神経細胞への入力とその出力に関わる複数のサブ回路の存在がより明らかになってきたことである[32]。これは前述の第三の根拠にも関連する。ドーパミン神経細胞を学習信号、さらには、何らかの誤差を利用した学習信号と捉えるとき、それを特定の報酬予測の誤差信号とするよりも、むしろ個々のサブ回路に適した誤差信号と考える方が自然である。そして、その複数のサブ回路が、感覚入力から行動出力までの予測と行動選択を修飾していると考えるならば、入力表現、予測学習、行動選択の学習が並列に進むと考えるほうがより適切である。

おわりに

与えられた紙面では残念ながら書ききれないが、報酬構造表現仮説の話題はまだまだ尽きない（なお、本著はテクニカルレポート[33]を元にしている)。たとえば、表現学習と予測学習が一緒に進むことでより能動的な予測が可能になる、ドーパミン神経細胞の投射先である神経細胞のシナプス可塑性の修飾だけでなく、それらの細胞活動の直接修飾が表現学習にも一役買っている、あるいは近年話題のモデルフリーとモデルベースそれぞれの予測と意思決定の区別にも新たな観点を与える、などである。ぜひ原著論文をご覧いただきたい[18]。報酬構造表現仮説は、予測と意思決定の脳機能解明に今後大きく貢献することが期待される。理論と実験の融合研究は、たとえばヒト fMRI 実験あるいは動物実験などで、今後ますます必要になってくるだろう（興味のある方は、NAKAHARA LAB: http://www.itn.brain.riken.jp をぜひご覧ください)。このような融合研究に挑戦する方々が増えていくことを、そして、それが価値意思決定や報酬予測の研究だけでなく、さらに多様な分野の研究にとっても大きな貢献をもたらすことを期待して

いる。

文献

1. Hikosaka, O., K. Nakamura, and H. Nakahara: Journal of Neurophysiology **95**(2): 567-584, 2006
2. Montague, P.R., B. King-Casas, and J.D. Cohen: Annual Review of Neuroscience **29**: 417-448, 2006
3. Rangel, A., C. Camerer, and P.R. Montague: Nature Reviews Neuroscience **9**(7): 545-556, 2008
4. Schultz, W.: Journal of Neurophysiology **80**: 1-27, 1998
5. Schultz, W., P. Dayan, and P.R. Montague: Science **275**(5306): 1593-1599, 1997
6. Sutton, R. and A.G. Barto: *Reinforcement Learning: An Introduction*. Adaptive Computation and Machine Learning series. 1998.
7. Barto, A., *Adaptive Critics and the Basal Ganglia*, in *Models of Information Processing in the Basal Ganglia*, J.C. Houk, J.L. Davis, and D.G. Beiser, Editors. 1994, pp. 12-31.
8. Houk, J.C., J.L. Adams, and A. Barto, *A Model of How the Basal Ganglia Generate and Use Neural Signals That Predict Reinforcement*, in *Models of Information Processing in the Basal Ganglia*, J.C. Houk, J.L. Davis, and D.G. Beiser, Editors. 1994, pp. 249-252.
9. Montague, P., P. Dayan, and T. Sejnowski: The Journal of Neuroscience **16**(5): 1936-1947, 1996
10. 中原裕之, *意思決定とその学習理論*, in *脳の計算論*, 甘利俊一 and 深井朋樹, Editors. 東京大学出版会. 2009, pp. 159-221.
11. Nakahara, H., H. Itoh, R. Kawagoe, et al.: Neuron **41**: 269–280, 2004
12. 中原裕之. *脳が意思決定をするとき*. RIKEN NEWS 2013 [380; 2-5]. 入手先: http://www.itn.brain.riken.jp/pdf/RN2013_02.pdf.
13. 独立行政法人理化学研究所. *記憶を使った脳の報酬予測のメカニズムの一端を解明*. プレスリリース 2004 Jan/22; 入手先: http://www.riken.jp/~/media/riken/pr/press/2004/20040122_1/20040122_1.pdf.
14. Bayer, H. and P. Glimcher: Neuron **47**(1): 129-141, 2005
15. Bromberg-Martin, E.S., M. Matsumoto, H. Nakahara, et al.: Neuron **67**(3): 499-510, 2010
16. Enomoto, K., N. Matsumoto, S. Nakai, et al.: Proceedings of the National Academy of Sciences of the United States of America, 2011

17. Satoh, T., S. Nakai, T. Sato, et al.: The Journal of neuroscience : the official journal of the Society for Neuroscience **23**(30): 9913-9923, 2003
18. Nakahara, H. and O. Hikosaka: Neuroscience Research **74**(3-4): 177-183, 2012
19. Schultz, W.: Current Opinion in Neurobiology: 10, 2012
20. Hayden, B.Y., J.M. Pearson, and M.L. Platt: Nature Neuroscience, 2011
21. Kolling, N., T.E. Behrens, R.B. Mars, et al.: Science **336**(6077): 95-98, 2012
22. Bao, S., V.T. Chan, and M.M. Merzenich: Nature **412**(6842): 79-83, 2001
23. Seitz, A.R. and H.R. Dinse: Current opinion in neurobiology **17**(2): 148-153, 2007
24. Zacks, J.M., C.A. Kurby, M.L. Eisenberg, et al.: Journal of cognitive neuroscience **23**(12): 4057-4066, 2011
25. Fiorillo, C.D., P.N. Tobler, and J. Schultz: Science **299**: 1898-1902, 2003
26. Bromberg-Martin, E.S. and O. Hikosaka: Neuron **63**(1): 119-126, 2009
27. Matsumoto, M. and O. Hikosaka: Nature **459**(7248): 837-841, 2009
28. Costa, R.M.: Current Opinion in Neurobiology, 2011
29. Daw, N.D., Y. Niv, and P. Dayan: Nature Neuroscience **8**(12): 1704-1711, 2005
30. Morris, G., A. Nevet, D. Arkadir, et al.: Nature Neuroscience **9**(8): 1057-1063, 2006
31. Roesch, M.R., D.J. Calu, and G. Schoenbaum: Nature Neuroscience **10**(12): 1615-1624, 2007
32. Lammel, S., A. Hetzel, O. Häckel, et al.: Neuron **57**(5): 760-773, 2008
33. 中原裕之, *報酬構造とその表現の学習―ドーパミン細胞をめぐる新仮説―*, in *BSI-ITN Tech Report No. 13-01* 2013.

図 1

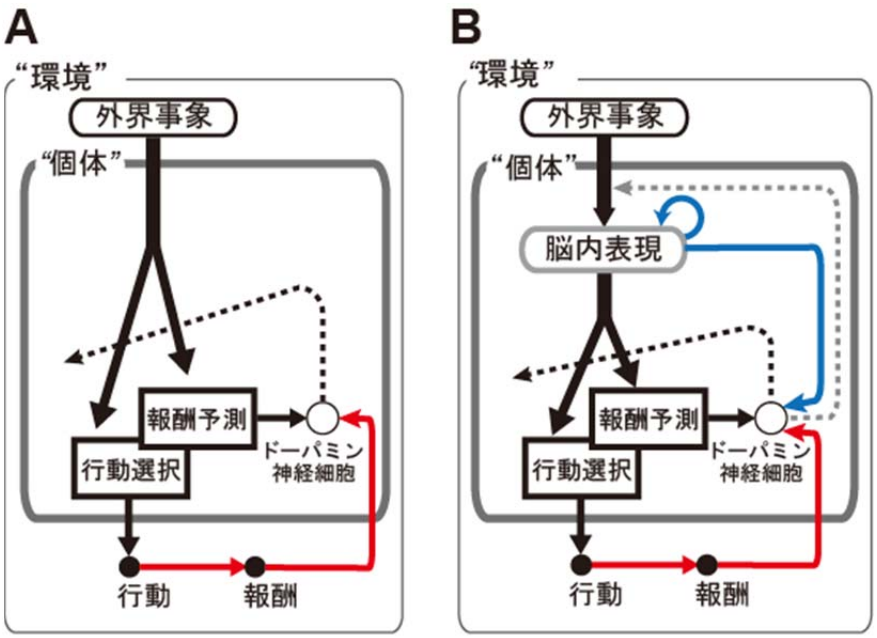

図 1　報酬予測誤差仮説（A）と報酬構造学習仮説（B）

（A）ドーパミン神経細胞活動は報酬予測誤差を表す（黒色点線矢印）。誤差は実際に得た報酬（ドーパミン神経細胞に入っていく赤色矢印）と予測された報酬との差である。ここで用いられている報酬予測は、基本的にはその各時点での外界事象（ドーパミン神経細胞に入っていく黒色矢印）で表現される予測に限られる。ドーパミン神経細胞活動で表される報酬予測誤差は学習信号として、報酬予測と行動選択の学習に寄与する（黒色点線矢印：本文中の学習で変化する「重み」が交差している黒色実線矢印に対応）。報酬予測誤差仮説では、ドーパミンの活動は特異的な報酬予測誤差シグナルを符号化すると考えられている。
（B）ドーパミン神経細胞活動は、報酬構造を学習するための信号を表す（黒色点線矢印および灰色点線矢印）。ドーパミン神経細胞は、報酬予測のみならず、学習された報酬構造に関する入力（青色矢印）を受ける。さらに、ここでの報酬予測は、各時点での外界事象と学習された報酬構造の両者を反映した脳内入力から生成される。この予測は、原理的に（A）で用いられた報酬予測より優れている。この予測を利用した報酬予測誤差信号はより優れた学習信号として、報酬構造学習の信号の一部となる（黒色点線矢印）。報酬構造成分をより多く含むドーパミン神経細胞活動は、報酬構造を反映した内的表現の学習にも用いられる（灰色点線矢印）。文献[18]の図を元に改変。