第 3 章　動機づけにおける予測

この章は書きにくい。長い中断状態だった。理由はこの領域は新しく、かつ研究が盛んで知見が日々変化しているからである。また、個人的な事情を述べると、そしてこちらの理由の方が圧倒的に強いのだが、わたしの世代は行動主義の洗礼を受けた世代である。その後認知主義が盛んになり、それまでの知識がグラついてきていた。加えて、専門が変わり、新しい知識を学んでこなかった。それでも、古い行動主義は、あやふやになりながらも、まだわたしの中に生き残っている。最近の強化学習（reinforcement learning, RL）などの論文を読むと、Thorndike や Tolman らがでてきてびっくりする。そして、あやふやな昔の行動主義の知識が顔を出し、うまく折り合いがつけられなくなるのだ。そういう事情があるが、ここでは予測という観点からまとめてみる（その点、他の章と比べると、骨と皮のような論文になってしまった）。いずれにせよ、報酬（強化）は学習、適応の原動力である。以下に述べるように、そこに「予測」が関係することは、「予測」がヒトを含めた動物の行動の基本的な特性であることを示している。

１．報酬予測誤差と中脳ドーパミン細胞

一般に条件刺激（conditioned stimulus, CS）と無条件刺激（unconditioned stimulus, US）を対提示する古典的条件づけ（classical conditioning, パヴロフ型条件づけ Pavlovian conditioning）において、US（強化刺激、報酬）に対するサルの腹側被蓋野 ventral tegmental area や黒質緻密部 substantia nigra pars compacta のドーパミン細胞は次のようなふるまいをする（Schultz et al., 1997）。すなわち、条件づけ（学習）の初期には報酬に対して一過的 phasic に反応するが、学習が進むにつれて US への反応は減少し、CS に一過的に反応するようになる。学習が完成すると、US への反応はなくなり、CS への反応が残る。これはドーパミン細胞の活動が報酬予測誤差（reward prediction error, RPE）、すなわち、予期した報酬と実際に得た報酬の差を反映するからと考えられた。学習初期には US は予期せぬもの、surprising であり、ドーパミン細胞は反応する（positive prediction error）。しかし、学習が進むにつれて CS が US の生起を予測するようになり、surprising でなくなり、報酬予測誤差が減少しドーパミン細胞は反応しなくなる。なお、予期された報酬がないと、ドーパミン細胞の反応は抑制される（negative prediction error）。

この考えは古典的条件づけの Rescorla-Wagner の理論に対応する（Rescorla & Wagner, 1972）。すなわち、かれらの理論は以下に定式化されている。学習の増加分 ΔV は

$$\Delta V = \alpha\beta(\lambda - V)$$

V は現在の連合強度、λ は連合強度の最大値、α, β はそれぞれ CS, US の salience を示す定

数である。λ-V が RPE に対応し、予期しない報酬が連合強度を強める。

Rescorla & Wagner の理論は無論ドーパミン細胞の振る舞いを説明するために考えられたものではない。古典的条件づけにおける現象を説明しようとした。その例として挙げられるのは阻止現象（blocking）である。この条件づけでは、条件刺激 A と報酬を対提示し、十分に条件づけを形成しておく（A+)。その後、新しい刺激 X を加えた AX を条件刺激として、条件づけを重ねる（AX+)。その後、刺激 X を単独で提示し、条件反応の出現を検討する。その結果、条件反応は見られなかった。すなわち、条件刺激 AX において、刺激 A があるので、報酬は十分に予測可能である。したがって、刺激 X は条件刺激として反応を惹起する力を持ちえないと解釈された。刺激 X は US と接近して対にされたが、条件づけが形成されるわけではない。また、条件制止（conditioned inhibition）という現象がある。条件刺激 A, B をそれぞれ単独で報酬と対にし、条件づけを形成する（A+, B+)。条件刺激 AX には無報酬で訓練をする（AX−)。その後、条件刺激 BX に対する反応をみると、反応は見られない。刺激 X は条件制止子であり、無報酬を予告すると考えられた。

では、ドーパミン細胞は阻止や条件制止でどのような反応を見せるのだろうか。それぞれ Waelti et al. (2001), Tobler et al. (2003) がサルで検討しているが、ドーパミン細胞の活動は行動の結果と一致した。Waelti et al. (2001) は A+, AX+の後に X を単独提示した結果、ドーパミン細胞は反応しなかった。上記のように、刺激 X は報酬に関して新たな情報を付加しないからと考えられる。さらに、B−, BY+の後に Y を単独提示したが、ドーパミン細胞は活動した。刺激 Y が報酬を予測するためと考えられる。Tobler et al. (2003) は A+, AX−の後に X の単独提示、B−, BY−の後に Y の単独提示をしたがドーパミン細胞は反応しない。その後、C+で条件形成をし、CX, CY への反応をみると、ドーパミン細胞は CX には反応せず、CY には反応した。刺激 X は条件制止子だが、刺激 Y は報酬がないことを予測する機能を獲得しなかったためと思われる。

このようなドーパミン細胞の活動は強化学習（reinforcement learning, RL）の理論では時間差（temporal difference, TD）誤差として、次の式で表されている。

$$\delta(t) = r(t) + \gamma V(t) - V(t\text{-}1)$$

ここで、r(t) は時間 t における報酬、V(t), V(t-1) はそれぞれ時間 t, t-1 における報酬の期待で、TD 誤差と呼ばれる。γ は時間割引要因。重要なのは、ドーパミン細胞は報酬そのものではなく、報酬の予測の誤差に応答していることである。TD 誤差理論は学習の進行によるドーパミン細胞の活動変化を次のように説明する。銅谷（2006）が解説していたので、それを紹介する。学習前は V(t), V(t-1) は 0 なので、δ (t) = r(t) で報酬に反応する。学習後には、報酬を予測させる刺激（CS）が提示されると、報酬期待の時間差分である γ V(t) – V(t-1) の応答がみられる。報酬が与えられた時点では、報酬 r(t) による正の成分と、もう報酬は期待できないという負の成分が打ち消し合い r(t) = 0 となり、ドーパミン細胞は反応しない。

一般に報酬、強化刺激は学習、すなわち永続的な行動変容、の最も重要な要素と考えら

れている。重要なことは、実質的な報酬が報酬そのものでなく、報酬の予測誤差にあるという点である。下等な動物を除き、ヒトを含めて動物は一般に先を見ながら、すなわち、予測をしながら生きている。それは前の二つの章で述べた感覚・知覚、運動・行為だけでなく、動機づけ、学習においても同様である。

下の図 3-1 は古典的条件づけの様々な操作とドーパミン細胞の活動の関係を模式的に図示してある（Schultz, 2007）。なお、強化学習理論は古典的条件づけだけでなく、道具的条件づけにも適用されている。

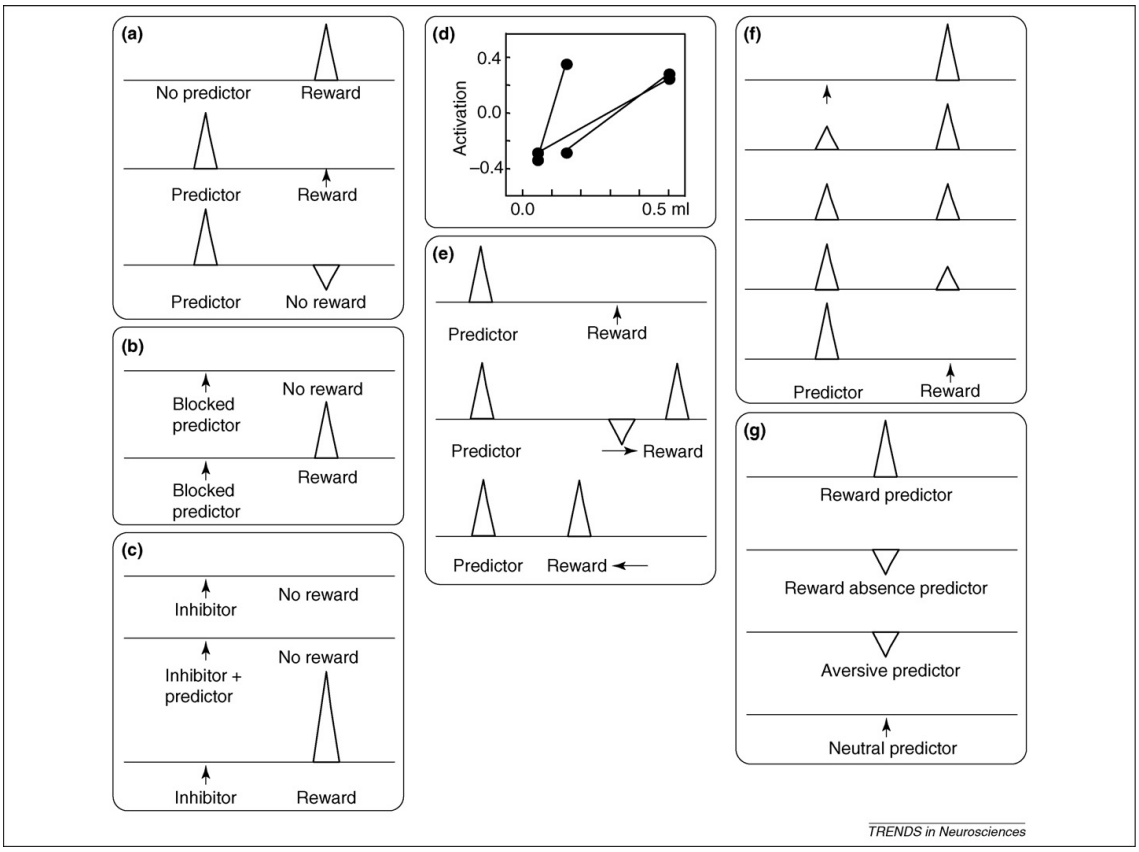


図 3-1. 条件づけの操作とドーパミン細胞の活動の関係。本文を参照して欲しいが、(a) は正負の予測誤差など、(b) は阻止現象、(c) は条件制止、(f) は学習過程を示している。予測を外れた報酬がドーパミン細胞を活性化させる。Schultz (2007) より。

2. 二種類の報酬予測（Pavlovian valuation）

古典的条件づけの重要性は条件刺激、弁別刺激（discriminative stimulus, S^D. 行動分析のタームで、道具的行動の自発のきっかけとなる刺激である。道具行動によって報酬がもたらされるので報酬とも結びつき、条件刺激と極めて類似する）と無条件刺激、強化刺激、

報酬また道具的条件づけにおける結果 outcome（これらのタームを区別せず使うことがある）を結びつけることである（S-O)。学習が進行するに従い、中脳のドーパミン細胞は US に対して反応しなくなり、報酬を予測する条件刺激、弁別刺激に反応するようになった。このような報酬予測誤差に対応した反応は中脳のドーパミン細胞以外でも見られるのだろうか。また、条件刺激、弁別刺激への脳の反応はすべて報酬予測誤差を反映したものなのだろうか。後者の疑問に関しては、報酬予測誤差に対応した活動と単に報酬を予測する活動の二種類があるようだ。

他の脳領域における報酬予測誤差の活動

中脳のドーパミン細胞のターゲット領域である線条体 striatum、特に報酬との関連が強い腹側線条体（側坐核、腹側被殻）ventral striatum (nucleus accumbens, ventral putamen) の細胞が報酬予測誤差に対応した活動を示すことはないようだ。しかし、古典的条件づけでラットの背側線条体 dorsal striatum にはドーパミン細胞と同じように報酬予測誤差をコードする細胞があった（Oyama et al., 2010)。また、やはり古典的条件づけで、サルの被殻の自発発射の高いニューロン tonically active neuron も予測誤差を反映した活性をした（Apicella et al., 2009)。ドーパミンそのものとの関係は、ラットの側坐核でボルタンメトリ voltammetry を使用して、ドーパミンを測定した Flagel et al. (2011) の研究がある（図 3-2)。条件刺激 CS に対して反応がでると無条件刺激 US への反応が減少し、報酬予測誤差

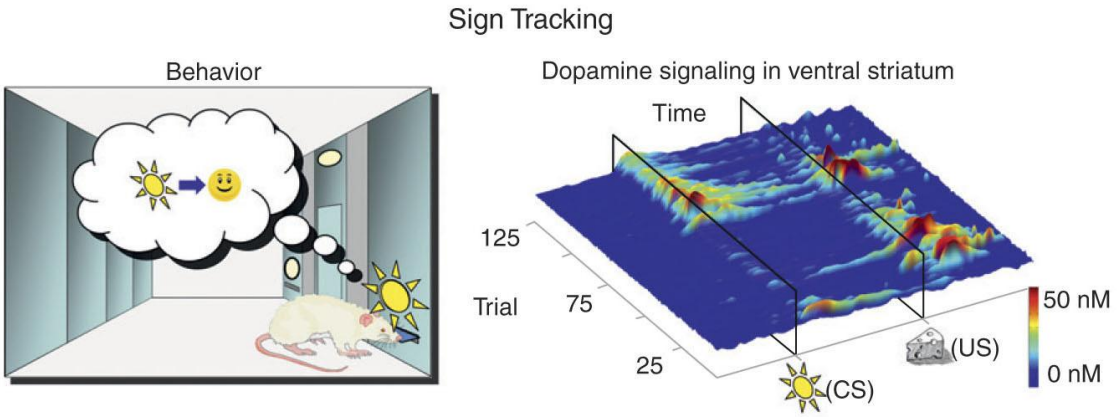


図 3-2 Clark et al. (2012) より。この図は Flagel et al. (2011) の研究に基づいている。sign tracking（後述）を示すラットは腹側線条体のドーパミンの動態は報酬予測誤差に対応。

に関係した反応がみられた。ヒトの fMRI 研究は、古典的条件づけと道具的条件づけを対比させている。古典的条件づけでは腹側線条体の腹側被殻が、道具的条件づけでは側坐核でドーパミン細胞と同じパターンの活性を報告している。道具的条件づけでは腹側線条体に予測誤差に対応した活性は見られず、背側線条体の前部尾状核 anterior caudate nucleus で予測誤差をコードする活性を見出した (O'Doherty et al., 2004; O'Doherty, 2004; 図 3-3)。

また、Hare et al. (2008) も一種の道具的条件づけで、腹側線条体に予測誤差に対応した活性を見出している。なお、O'Doherty et al. (2002) の研究では側坐核、扁桃核後部 posterior amygdala が US よりも CS で大きな活性を示した。腹側線条体に関して、fMRI では予測誤差に対応する活性が見られるのに、ニューロン活動ではそれが見られないこと、背側線条体では古典的条件づけで予測誤差に対応したニューロン活動が見られるのに、ヒトの fMRI ではそれがない。これらについてはさらに研究が必要だろう。

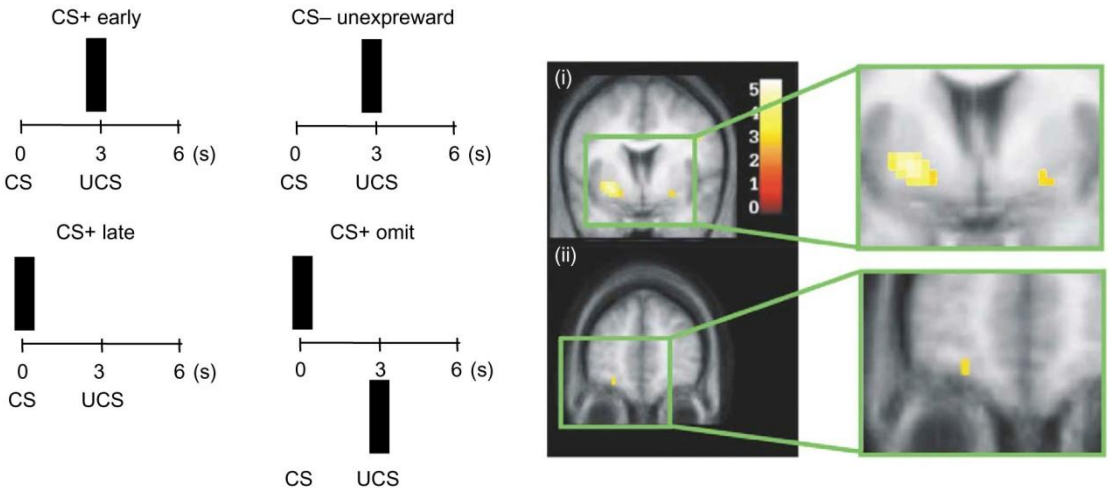


図 3-3. 左は時間差誤差説に基づく脳活性のパターンの模式図。右はそのパターンを示した腹側被殻（上）と眼窩前頭部皮質（下）。O'Doherty (2004) より。この図は O'Doherty et al. (2003) に基づく。

腹側線条体と並んで報酬や価値 value に関係するが眼窩前頭皮質 orbitofrontal cortex である（Rolls, 2000）。なお、腹内側前頭前皮質 ventromedial prefrontal cortex も含めることがある。松元健二（personal communication）によると、サルの眼窩前頭皮質のニューロンは課題中の報酬よりも予期せず与えられた報酬に強く反応したという。報酬予測誤差をコードしているのかもしれない。ヒトの fMRI 研究では図 3-3 に示すように、報酬予測誤差を反映している活性が見られた。

予測誤差を反映しない報酬予測の活動

では、CS や SD への応答はすべて報酬予測誤差を反映したものなのだろうか。腹側線条体のニューロンは、一過的なドーパミン細胞の発射と異なり、持続的 tonic な反応を示す（ラット、Day et al., 2006 : サル、Schultz et al., 1992）。ドーパミンの動態に関しては、図 3-4 は図 3-2 の続きである。腹側線条体のドーパミンは CS にも US にも反応している。報酬予測誤差とは異なる反応パターンである。興味深いことに図 3-2 と図 3-4 はラットの系統が異なる。それは行動にも現われるのだが、それは 4. model-based vs model-free reinforcement

learning に関係するので、そちらで詳しく述べる。ここで重要なのは、腹側線条体には、少なくとも、報酬予測誤差（図 3-2）と、条件づけの進行に伴い、CS と US の両方に反応し報酬予測の機能を持つと考えられる（図 3-4)、2 つのタイプのニューロンがあることである。

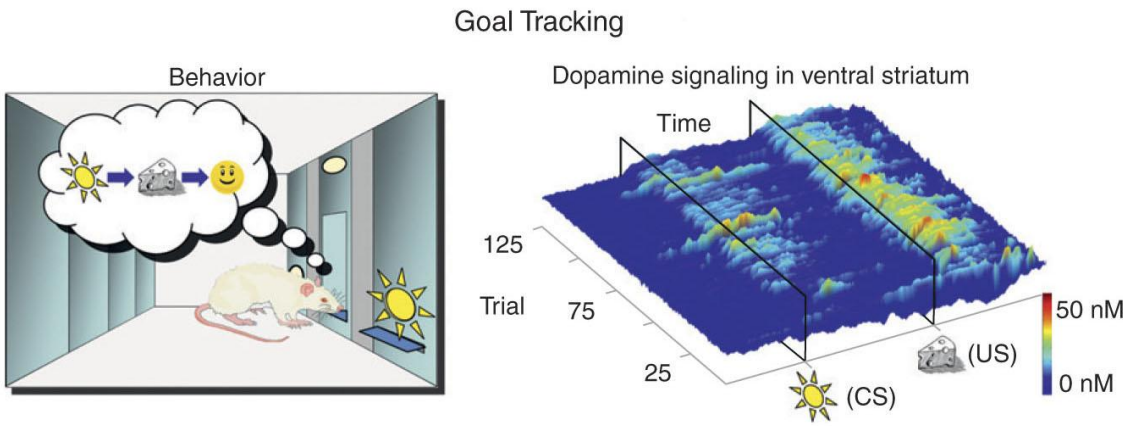


図 3-4. Clark et al. (2012) より。この図は Flagel et al. (2011) の研究に基づいている。goal tracking（後述）を示すラットは腹側線条体のドーパミンの動態は報酬、報酬の予測に対応。

サルの眼窩前頭部には様々な報酬に個別的に反応するニューロンがある（Rolls, 2000)。すなわち、報酬の identity に関係する。これらのニューロンは飽食によって活動が低下するので、報酬の感覚的側面よりは価値 value に関係すると考えられている（サル : Critchley & Rolls, 1996b, ヒト : O'Doherty et al., 2000; Valentin et al., 2007)。そして、これらのニューロンは、学習により、報酬を予告する刺激にも反応するようになる（例えば、Critchley & Rolls, 1996a; Tremblay & Schultz, 1999)。すなわち、図 3-4 と同様の反応パターンを示す。そして、報酬に対して個別的であったように、特定の報酬を予告する（Schultz et al., 1998; Tremblay & Schultz, 1999)。ヒトの fMRI 研究は、すでに述べたように、報酬予測誤差に対応する活性を示したが（図 3-3)、報酬価、その予測に対応する活性もある（Hare et al., 2008)。

これらの結果を予測という観点からみた場合、中脳のドーパミン細胞、腹側線条体、扁桃核、眼窩前頭部皮質とそれに連なる腹内側前頭前野等によって構成される脳内報酬系の重要な役割は、報酬予測誤差と報酬（価値）予測というように、予測にあることを示している。そして、両者は共同して意思決定に関与しているようだ（Daw et al., 2005; Gläscher et al., 2010; Daw et al., 2011)。

3. goal-directed vs habitual actions

道具的行動には目標指向的 goal-directed と習慣的 habitual な行動がある。ある行動を学習する場合、初めは目標指向的だが、すっかり学習すると習慣的になると考えられている。これらはそれぞれ、刺激 stimulus－反応 response－結果 outcome, S-R-O 学習と刺激－反応、S-R 学習と表現される。この二つは飽食や味覚嫌悪条件づけなどの価値低下 devaluation や随伴性低下 contingency degradation 操作により区別される。Adams (1982) はラットのレバー押しを蔗糖で訓練し、味覚嫌悪条件づけの影響を消去の形式でテストした。訓練途中の目標指向的なラットでは嫌悪条件づけの影響が見られたが、すっかり訓練した習慣的なラットではその影響は見られなかった。Dickinson et al. (1998) は十分に訓練が行われたラットは、omission スケジュールの導入による随伴性の変化に、訓練途中のラットと比較して、鋭敏でないことを示した。なお、ここではふれないが、この二つの行動については多くの研究があるようだ（たとえば、Kosaki & Dickinson, 2010 など）。

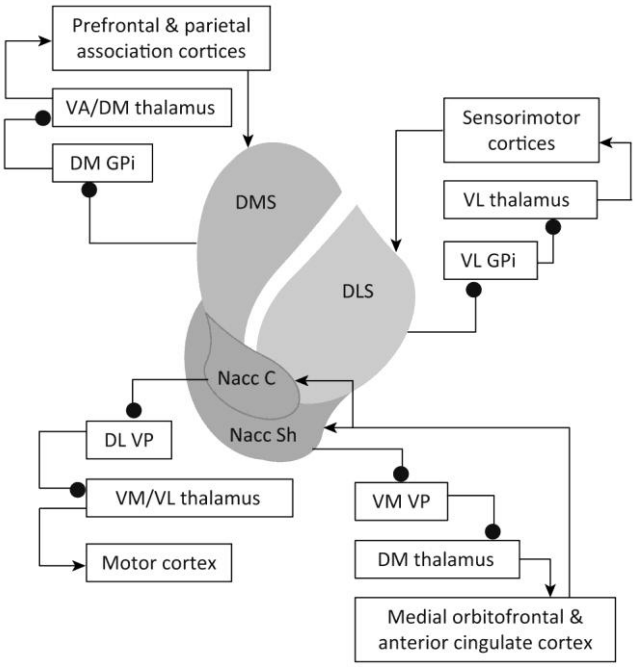


図 3-5. 線条体の区分と線維連絡。Liljeholm & O'Doherty, 2012 より。背側（薄い灰色）は内側 DMS と外側 DLS に分けられ、腹側の側坐核（濃い灰色）はコア Nacc C とシェル Nacc Sh に分けられる。

以上、ラットの研究だが、この二つの行動は関係する線条体の脳領域が異なる。線条体を背側と腹側に二分する。背側は主に道具的条件づけ instrumental conditioning に関係し、強化学習の actor/critic の理論（銅谷 (2006) に解説がある）では、actor の役割を果たすと考えられている。背側部は内側、外側に分けられ、内側が目標指向行動、外側が習慣的行動に関係する。内側、外側部は他の脳領域との結びつきが異なっており、内側部は前頭皮質 frontal cortex などの連合野と、外側部は感覚運動領野 sensorimotor cortex との結びつきが強い。したがって、内側部は認知的、外側部は感覚運動的な面があり、それぞれ目標指向的、習慣的行動の特性と合致している。ヒトではそれぞれ尾状核、被殻に対応すると考えられている。線条体の腹側部は古典的条件づけに関係し、動機づけ、報酬、結果、価

値、すなわち、critic の役割を果たすと考えられている。すなわち、Pavlovian valuation に関係する。腹側部にある側坐核はコア core とシェル shell に区分される。その機能から分るように、線条体腹側部は中脳の腹側被蓋野のドーパミン細胞、視床下部 hypothalamus、扁桃核、眼窩前頭部、前部帯状回 anterior cingulate gyrus などの大脳辺縁系や関連領域と密な連絡がある。図 3-5 を参照されたい。

4. model-based vs model-free reinforcement learning

ここで興味深いのは、目標指向的、習慣的行動の区分から発展した、2 種類の強化学習があるという考えである。目標指向行動に対応するのが model-based 強化学習, 習慣的行動に対応するのが model-free 強化学習である。2. で紹介した Flagel et al. (2011) の実験を説明する（図 3-2, 図 3-4)。Flagel らは CS の位置と US のでる位置を離した状況を設定した。その結果、行動が異なる 2 系統のラットがいた。一方は、CS が提示されると CS に接近し、CS 終了で US の方に移動した（図 3-2)。刺激が接近反応を引き起こしたと考えられる。もう一方は、CS が提示されると、US の方に移動し、そこで CS が終了するまで待っていた（図 3-4)。こちらでは刺激は報酬（US）の表象を引き起こしたと考えられる。前者が sign-tracking で model-free 強化学習に対応し、後者が goal-tracking で model-based 強化学習に対応する。

Doll et al. (2012) によると、前者の model-free 強化学習はすでに述べた報酬予測誤差を学習信号、教師信号とするものである。その特徴は報酬を最大にするために、学習者は課題の連続的な推移構造（world model）を理解する必要がない点にある。Model-free と称される所以である。どの行為が利益をもたらすかを、結果として生じる報酬に関する経験から直接的に学習する。その意味で、行動主義心理学的であり、retrospective な性格を持つ。この学習は計算の負荷が小さいのが利点だが、環境の変化に対応して行動を柔軟に変化させる必要がある状況では、対応が難しくなる。後者の目標指向行動に対応する model-based 強化学習は、ある課題における事象と行為の随伴性を学習する。すなわち、どの行為はどの結果をもたらすかとか、迷路のどの通路を通ればどこに行くかとかの学習である。そして、結果を内的に予測、シミュレートすることにより、理想的な行為を適応的に計算する。この学習は変化する環境に柔軟に適応できる。この world model（これは内部モデル、順モデル forward model である）の利用が model-free と異なる点である。その意味で、認知心理学的であり、prospective である。ただ、この学習は計算が膨大になる可能性があり、それを避ける手段、推定が必要となる。

実際の場面でこれら二つの学習のいずれが働いているかを検討した研究がある。Daw et al. (2011) の連続的二肢選択マルコフ決定課題 sequential two-choice Markov decision task の fMRI 研究がそれである（図 3-6 参照)。この課題では 2 つの刺激が提示され、いず

れかの刺激の選択が求められる。その結果、図 3-6 の上図 B にあるように、いずれの選択でも 7 対 3 の割合でそれぞれ異なる第 2 段階の刺激対の状態になる。そして、それぞれの状態で再び選択反応を行い、その結果、変動する報酬か無報酬となる。下図は第 1 段階の選択行動に関するもので、A, B はそれぞれ強化(model-free), model-based 強化学習からの選択行動の予測、C は実際の行動の結果である。C から分るように、両方の強化学習の要素が含まれているような結果になった。

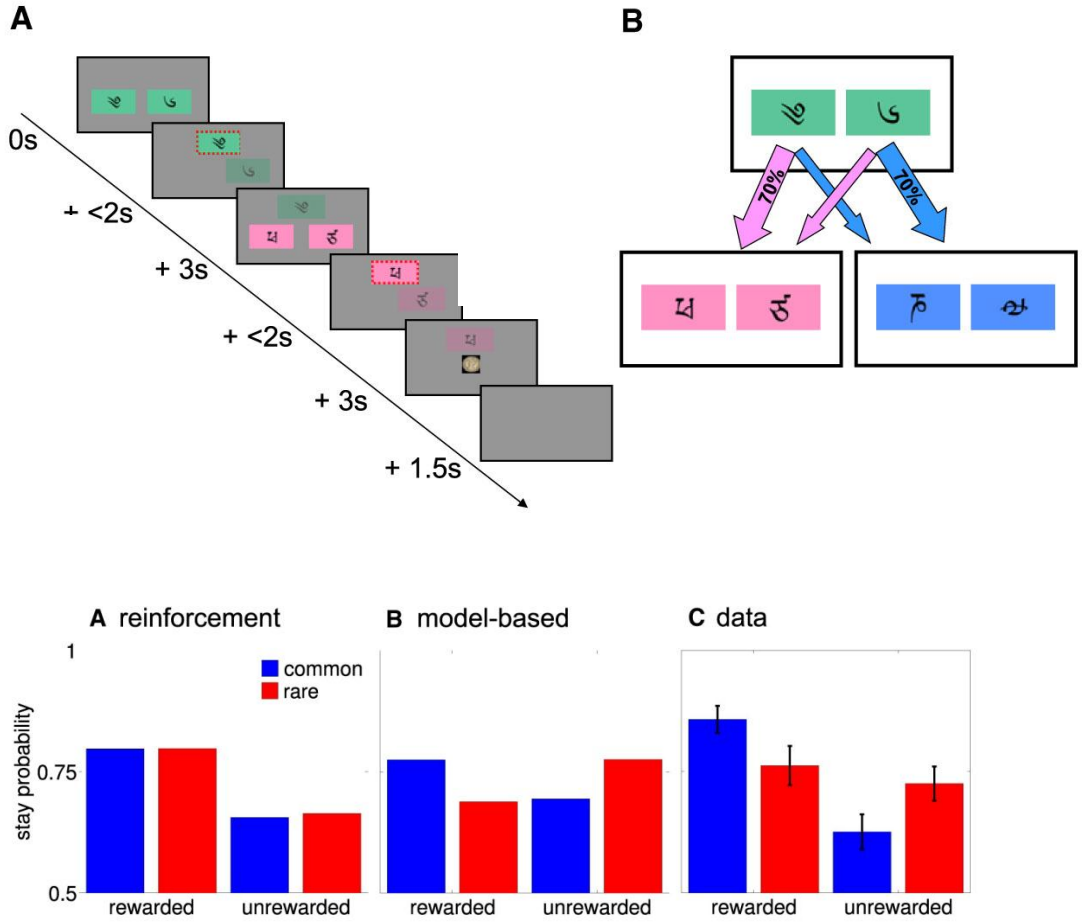
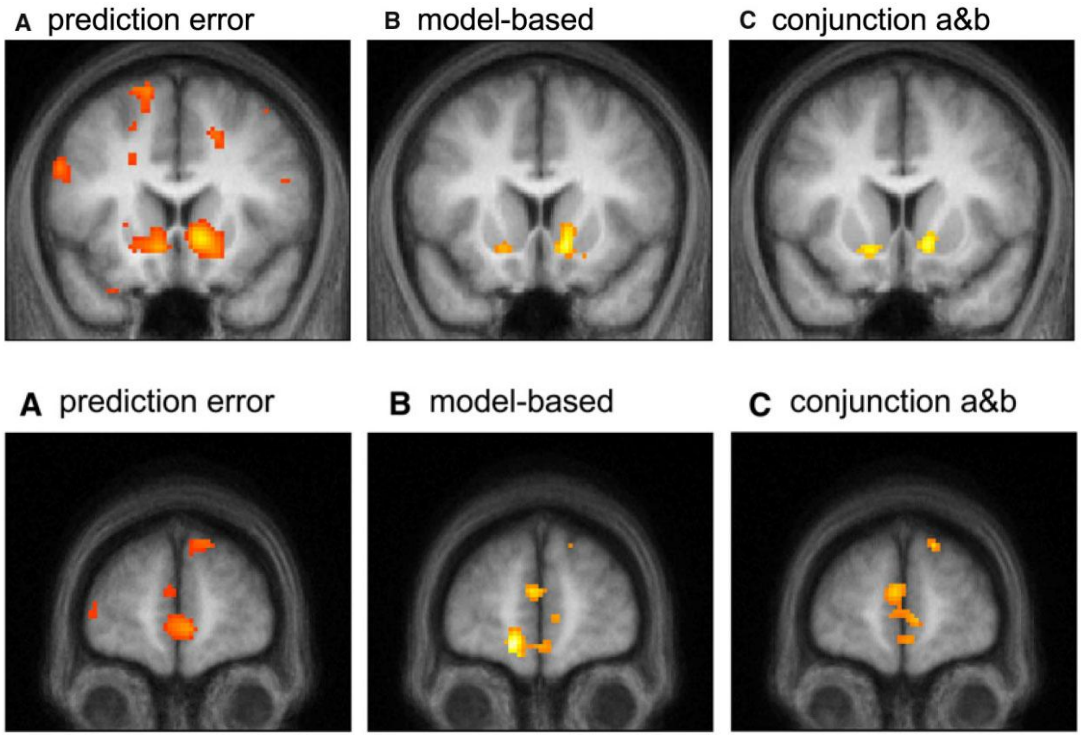


図 3-6. 上：A は課題の時間的な変化。B は 2 段階の選択課題の状態遷移の構造。下：第 1 段階の選択関するもので、A, B はそれぞれ強化 reinforcement (model-free), model-based 強化学習からの予測で、C は実際の結果。青は 70%, 赤は 30%の状態遷移。Daw et al. (2011) より。

fMRI による脳の活性は線条体と内側・腹内側前頭前野で検討した。model-free の報酬予測誤差、それに model-based の成分を考慮した場合（model-based の予測から model-free の報酬予測誤差を引いたもの）に分けて脳活性との相関を検討した。図 3-7 に結果を示す。図で上が線条体、下が内側前頭前野の結果である。A が model-free による報酬予測誤差、

B が model-based による予測を考慮した結果、C が両者の conjunction である。この結果は両脳領域において model-based と model-free 強化学習の両方が含まれていることを意味する（hybrid theory, Gläscher et al. 2010）。両者は個人や課題によって重みづけが変わるのだろう（図 3-7 の下段）。



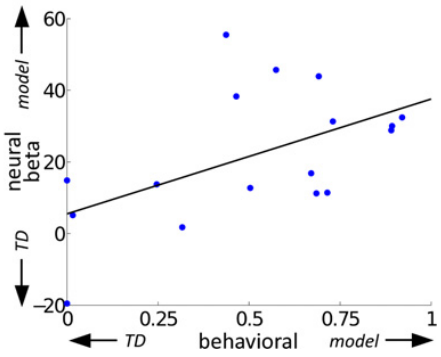


図 3-7. 上段は線条体、中段は内側前頭前夜の結果。A は model-free の報酬予測誤差との相関領域、B はそれに model-based の成分を加味した結果。C は A,B の conjunction. 下段は右の腹側線条体における報酬予測誤差－model-based の関係を行動と脳の活性で検討したもので、有意な正の相関がみられた。Daw et al. (2011) より。

5. 動機づけにおける予測

以上、動機づけにおける予測と脳の活動をまとめてみた。おそらく、十分に理解できていないところがあるだろうし、あまりにも図式的に単純化して解説したところもあると思う。その点はご容赦いただきたい。

Model-based の強化学習は prospective な内部モデル、順モデルを含んでいる。これは第 2 章の運動制御理論と同じ性質のものである。運動制御では運動の予測が問題となるが、model-based の強化学習では world model、すなわち、状態 state の構造やその遷移の予測が問題になる。そして、model-based, model-free 強化学習に共通して、報酬の予測が重要な役割を演じている。二種類の強化学習、二種類の報酬予測が同時に機能して、ヒトを含む動物の行動を適応的にしていると思われる。動物は単に刺激に反応する存在ではない。能動的に予測をしながら生きている。

引用文献

Adams, C.D. (1982) QJEP, 34B:77-98.

Apicella, P. et al. (2009) EJNS, 30:515-526.

Clark, J.J. et al. (2012) COINB, 22:1054-1061.

Critchley, H.D. & Rolls, E.T. (1996a) JNP, 75:1659-1672.

Critchley, H.D. & Rolls, E.T. (1996b) JNP, 75:1673-1686.

Daw, N.D. et al. (2005) NNS, 8:1704-1711.

Daw, N.D. et al. (2011) Neuron, 69:1204-1215.

Day, J.J. et al. (2006) EJNS, 23:1341-1351.

Dickinson, A. et al. (1998) QJEP, 51B:271-286.

Doll, B.B. et al. (2012) COINB, 22:1075-1081.

銅谷賢治（2006）数理科学, No.512:1-8.

Flagel, S.B. et al. (2011) Nature, 469:53-57.

Gläscher, J. et al. (2010) Neuron, 66:585-595.

Hare, T.A. et al. (2008) JNS, 28:5623-5630.

Kosaki, Y. & Dickinson, A. (2010) JEP:ABP, 36:334-342.

Liljeholm, M. & O'Doherty, J.P. (2012) TICS, 16:467-475.

O'Doherty, J.P. (2004) COINB, 14:769-776.

O'Doherty, J.P. et al. (2000) NRep. 11:893-897.

O'Doherty, J.P. et al. (2002) Neuron, 33:815-826.

O'Doherty, J.P. et al. (2003) Neuron, 28:329-337.

O'Doherty, J.P. et al. (2004) Science, 304:452-454.

Oyama, K. et al. (2010) JNS, 30:11447-11457.

Rescorla, R.A. & Wagner, A.R. (1972) A theory of Pavlovian conditioning: variations in the effectiveness of reinforcement and nonreinforcement. In Black, A.H. & Prokasy, W.F. (eds.) Classical Conditioning II: Current Research and Theory, pp.64-99, New York: Appleton Century Crofts.

Rolls,E.T. (2000) CC, 10:284-294.

Schultz, W. (2007) TINS, 30:203-210.

Schultz, W. et al. (1992) JNS, 12:4595-4610.

Schultz, W. et al. (1997) Science, 275:1593-1599.

Tobler, P.N. et al. (2003) JNS, 23:10402-10410.

Tremblay, L. & Schultz, W. (1999) Nature, 704-708.

Valentin, V.V. et al. (2007) JNS, 27:4019-4026.

Waelti, P. et al. (2001) Nature, 412:43-48.

雑誌の略称

CC: Cerebral Cortex

COINB: Current Opinion in Neurobiology

EJNS: European Journal of Neuroscience

JEP:ABP: Journal of Experimental Psychology: Animal Behavior Processes

JNP: Journal of Neurophysiology

JNS: Journal of Neuroscience

NNS: Nature Neuroscience

NRep: Neuroreport

QJEP: Quarterly Journal of Experimental Psychology

TICS: Trends in Cognitive Sciences

TINS: Trends in Neurosciences