題目：脳の計算理論：強化学習と価値に基づく意思決定

中原裕之（なかはら　ひろゆき）
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〕

はじめに

意思決定の脳機能を理解するには、意思決定時に行われる「予測」の役割と、その予測が「学習」されるプロセスを解明することが本質的な課題となる。意思決定には行動の選択が伴い、その選択は未来の予測に基づいて行われる。行動選択とは、その選択した行動がもたらすであろう未来を選択することに他ならない。予測を学習できる、つまり、経験に基づいて予測を適応的に変更できることが、適切かつ柔軟な意思決定につながる。近年、報酬予測（＝価値）に基づく意思決定（value-based decision-making; 以下、「価値意思決定」）と予測学習における脳機能の解明は著しい発展を遂げてきた[1,2]。これには、実験研究と理論研究の相互交流、なかでも強化学習（reinforcement learning）[3]の計算理論を用いた実験と理論の融合研究が大きく寄与している[4]。本稿では、この強化学習を中心に、価値意思決定に関する脳の計算理論を概観する。

強化学習の理論を土台にすることで価値意思決定の計算プロセスが明確化し、それが行動と神経活動の対応づけを可能にする。これにより、脳の価値意思決定の規範的理解（normative understanding）が得られ、脳活動の目的と意味が理解できるようになる[5,6]。強化学習に限らず、計算理論を土台にする研究の重要性は近年ますます注目されている。その根本的理由は、脳計算の具体性である[5]。脳計算というと、ややもすると抽象的で実体を欠いた概念と思われがちだが、それはまったくの誤解である。脳計算の持つ具体性こそが、意思決定などの複雑な脳機能解明に明快な展望を与えてくれる。以下、次節で強化学習を概説した上で、この分野の土台であるドーパミン神経細胞活動の報酬予測誤差仮説に触れる。さらに、外界の脳内モデルを用いる価値意思決定（典型例として model-based reinforcement learning; 以下、「モデルベース強化学習」）を紹介する。その後、他人の心の内部モデルを利用する意思決定について述べ、最後にドーパミン神経細胞に関わる新仮説である、「報酬構造学習仮説」に触れる。

次節を始める前に、関連する脳部位を簡単にまとめておく。価値意思決定に関わる脳領野は多岐にわたる。その多くが中脳ドーパミン神経細胞の投射を受けている。それらの領野は、知覚から運動、あるいは、報酬予測から行動選択まで、さまざまな段階に関与する。領野の機能的分化は、知覚から運動の軸、そして以下で議論する、異なる内部モデルの利用という軸、この 2 つの軸に沿っていると思われる。意思決定に関わる主た

る皮質下領域には大脳基底核（すなわち線条体、側座核、黒質網様部、淡蒼球)、 中脳ドーパミン領域（黒質緻密部、腹側被蓋野）および扁桃体が含まれる [7)]。その他にも、報酬の欠如と嫌悪刺激に関連する手網外側核、報酬価値に関連する視床下部、運動出力に関連する上丘と脚橋被蓋核、およびセロトニンニューロンを含む背側縫線核がある。海馬は種々の内部モデルの構築を支援する [8)]。主たる皮質領域は前頭前皮質および頭頂皮質、特に背外側前頭前皮質、内側前頭皮質（前帯状皮質を含む）、眼窩前頭皮質である [9-11)]。特に腹側内側部前頭前皮質は最終的に選択される価値をコードすると思われる。また、この領域は他の社会的神経回路、たとえば（後部）上側頭溝や側頭頭頂接合部とともに社会的意思決定にも関与する。

強化学習：理論的背景とモデルフリー強化学習

強化学習の理論的背景を簡単にまとめた上で、理論の基本となる「モデルフリー強化学習」を紹介する。「価値」と「報酬予測誤差」は強化学習における重要な変数である。モデルフリー強化学習について、まずは逐次的意思決定が関与しない場合から理解しよう（図 1A)。「エージェント（意思決定者)」は所与の環境において「状態」を観測し、選択肢の中から報酬を得るための「行動」を選択する。価値とは、この行動選択をするための予測報酬であり、報酬予測誤差（報酬予測誤差＝実報酬－予測報酬）は将来の予測（および行動選択）を改善するための学習信号として機能する。報酬予測の学習は誤差の減少を目指す（新たな報酬予測＝今までの報酬予測＋学習係数×報酬予測誤差)。これが「モデルフリーの強化学習」と呼ばれるのは、その意思決定（報酬予測と行動選択）に環境の内的モデルを一切必要としないことに由来する（図 3、緑色矢印)。報酬予測と行動選択は状態からの直接連想として学習され、またこの 2 つのみが意思決定に用いられるのである。さて、この図式をより完成されたものにするには、「時間」の観点を含めることが必要で、そこで強化学習理論の中で最もよく用いられる時間差学習（temporal difference learning; 以下 TD 学習）が出てくる。たとえば短い時間スケールでは、いわゆる実時間の視点、すなわち上の例で言えば、状態の観測から報酬獲得までの時間経過が考えられる。TD 学習は、この実時間を適切に扱うのである（これが次節で述べる報酬予測誤差仮説の提唱につながったのである)。より長いスケールでは、それが逐次的意思決定につながる。ある状態で行動選択をしてすぐに報酬を得るというよりも、いくつかの状態を経由し（その各状態で各行動を選択し)、その経過の途中や最後に得られる報酬を集める。このとき、最大限の報酬を獲得するためには、エージェントは即時の報酬を追求するだけではなく、即時の報酬と将来の報酬のバランスも確保しなければならない。TD 学習はこれらの問題に対処し、得られる報酬のバランスを状態間、つまり時間経過で考慮した価値を形成する。その学習には TD 誤差が用いられる。

ドーパミン報酬予測誤差仮説

価値意思決定の研究が著しく発展したきっかけは、ドーパミン報酬予測誤差仮説の提唱であった[4)]。この仮説は、ドーパミン神経細胞活動の実験結果に対して、モデルフリー強化学習（より正確には、TD 学習）理論と 2 つの仮定を土台に提唱された。この仮説の骨子は、ドーパミン神経細胞活動が「TD 学習の報酬予測誤差（TD 誤差）を表し、それを学習信号として、報酬予測が学習される」というものである（図 1B、図 2）。この仮説に用いられた 2 つの仮定を整理すると、第一の仮定は、最適の意思決定を下すためには現在の「状態」が入力としてあれば十分で、過去の情報は不要というものである。この仮定は、現実世界では破られる可能性はあるが、理論的な解析を容易にする（後述）。第二の仮定は直近の感覚入力のみが「状態」を形成するという仮定である。これは、単純化するための仮定である。実際、報酬予測と行動選択の両方に用いられた「状態」を実験条件で明示的に与えることはできるが、脳の強化学習に用いられる脳活動としての入力（つまり「状態」）として特定するのは困難なことが多い。

内部モデルを用いる意思決定：モデルベース強化学習

ドーパミン報酬予測誤差仮説の提唱とともに導入されたモデルフリー強化学習の理論は、その後、多くの価値意思決定に関する脳研究に用いられ、大きな成果を挙げた。しかし、感覚入力からの直接連想だけで、報酬予測と行動選択がいつも決まってしまうわけではない（図 3）。生物は経験を通じて、外界の事柄をいわば脳内シミュレーションするような内部モデルを学習することができる。もちろん、内部モデルを外界の環境すべてについて獲得しているわけではないが（それは極めて困難である）、環境構造のさまざまな側面について種々の内部モデルを獲得していると考えられる。特に大切な内部モデルとしては、報酬にまつわる環境構造や、社会生活における他者の心などが挙げられる。現在、それらの内部モデルを含む価値意思決定の研究が花開きつつある。その一例として、索餌行動（foraging）の研究がある。この行動では、報酬をもらったからそのまま報酬予測を上げて、その場所に留まり引き続き餌を探すのか、もうそこを離れて未知の場所に餌を探しに行くのかを意思決定しなければならない。留まるか離れるか、どちらの行動の価値を考えるにも内部モデルが必要である。他の例としては、モデルベース強化学習の研究も内部モデルがその土台となっている[8,12)]。一般にモデルベース強化学習では、「状態から状態への遷移」（状態遷移）および「状態と行動のペアに対して起きる報酬」（報酬関数）のいずれか一方あるいは両方の内部モデルを（たとえ近似的にでも）学習する。そして、それらの学習をもとに、報酬予測と行動選択の意思決定を行う（図 3、青色矢印）。このモデルベース強化学習の特性を生かした研究としては、たとえば、目的指向性の行動と習慣的な行動の間の差を研究するために、モデルベースド

強化学習のほうが意思決定に柔軟性があることを利用する研究がある[13]。その他に、学習時の特性について、あるいはモデルフリー強化学習とモデルベース強化学習がどのように協調的に働くかなどについても研究が進められている[8,11,14,15]。

他者の心の学習

「相手の気持ちを考えなければ…」――私たちは生涯で何度この言葉を繰り返すことだろうか。私たちは他人の心について内部モデルを持っている。これがよく知られている「心の理論」（theory of mind）である[16]。実は、強化学習の脳計算理論は、価値意思決定を超えて、社会的知性の脳機能の解明にも役立つことが示されつつある。これは、「はじめに」で述べたように、脳計算の具体性を追求することで、社会的知性という複雑な脳機能が、明快な強化学習の脳計算理論を通じて明らかにされるからである。この分野は今後の発展が大いに期待されている。

ここで私たちの研究例を挙げて説明したい[17]（図 4）。心の理論の土台として、「人は自分の心のプロセスをもとにして、他人の心のプロセスを自分の心の中で構成する」というシミュレーション説（simulation theory）がある[18]。一方、「シミュレーションは不必要で、人は他人が何にどう反応するかのパターンを学習して、他人の目に見える行動を当てる」という考え方もある（例「セオリー・セオリー」（theory-theory）；以下、「行動パターン説」と総称）[19]。これらの議論を踏まえて、私たちは、他者の心の動きを予測するという複雑なプロセスを解明するには、その意思決定プロセスと学習について徹底的な検証を行うこと、すなわち脳情報処理、つまり脳計算の観点から検討することが必要だと考え、他者の心の脳内シミュレーションに迫るために、価値意思決定の実験を行った。その際の工夫として、被験者に価値意思決定の課題（コントール課題）を行わせると同時に、他人がコントロール課題をしているときの、その他人の行動選択を予測するという課題（メイン課題）も行わせた。そして、モデル化解析手法と呼ばれる脳計算モデルを利用して、行動データと脳活動を照合し、脳計算過程に対応する脳活動をヒト fMRI 実験で解析した。その結果、実は、ヒトは他者の心をシミュレーションするとき、2 つの学習信号を用いることが分かった。一つは、まさしくシミュレーション説が唱える他者報酬予測誤差信号（simulated-other’s reward prediction error）で、これは前頭葉腹内側部にのみ対応する脳活動が見られた。一方、その学習信号だけではなく、行動パターン説に当てはまるような、他者の行動の予測と実際の行動との差、つまり他者行動予測誤差信号（simulated-other’s action prediction error）を利用した学習も同時に行われることがわかった。他者行動予測誤差は、前頭葉背外側部や背内側部の脳活動に表れていた。他にも、社会性の議論でしばしば挙げられる脳領野、たとえば、側頭頭頂接合部と後部上側頭溝でも対応する脳活動が見られた[20,21]。私たちは、強化学習に基づき脳計算の具体性を追求することで、ヒトは 2 つの説を統合したハイブリッドな学習を通じ

て、他者の心（価値意思決定）をシミュレーションしていることを発見した。

より広がる強化学習：ドーパミン報酬構造学習仮説

ドーパミン神経活動は、感覚入力を状態としてモデルフリーの学習信号（TD 誤差）だけをコードしているのだろうか。実は、私たちは実験と理論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[22)]。紙面の限りから、喩えを使って説明しよう。たとえば、報酬予測誤差仮説は、「報酬」を「日曜」に置き換えて考えた場合、「明日は日曜ですか」という質問の場合、「はい」という答えがくる確率は（1 週間は 7 日あるため）7 分の 1 であるという予測が行われるとしている。しかし、私たちが得た結果では、「今日が土曜日なら確率は 1 である」という予測が働くことが示された。また、報酬予測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が、誤差仮説に触発されて積み上げられた近年の知見は、より多様な情報（例：不確実性、事前情報、運動開始、アラート信号、サリエンス信号）がその活動を修飾していることを示している[23) 24)]。しかも特筆すべきは、これらはすべて原理的に脳内表現の学習を助ける信号になりうる点である。また、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が支持されている[25)]。

これらの知見をもとに、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 5）[26)]。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に留まらず、環境構造も反映する学習信号であり、それは予測の学習だけでは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考えられる。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双方に影響を与えるとするのがこの仮説である。この議論のカギは、脳の価値意思決定における「状態」とは何かを理解することである。今まで、この疑問はほとんど見過ごされてきた。しかし、モデルフリー強化学習を行う脳回路への入力が「状態」であるはずだから、感覚入力以外の内的に生成された情報も脳内表現の一部であって不思議はない。報酬構造の内部モデルを利用することによって、より優れた脳内表現が作られる（図 3、脳内表現と内部モデルの間の赤色矢印）[26)]。このモデルフリー強化学習は報酬予測誤差仮説のモデルフリー強化学習に比べて強力である。このモデルフリーの意思決定は、内部モデルの利用という意味で実はモデルベースでもある。それでも、学習と意思決定自体は、「状態」からの直接連想という意味ではモデルフリーである。この直接連想では、より正確な報酬予測と行動選択が生成されうるし、その学習は、ドーパミン活動がそれを反映した報酬予測誤

差をコードすることで可能になる [26,27]。

さいごに

価値に基づく意思決定について、強化学習を中心とした脳の計算理論を概観し、その最近のトピックを駆け足で見てきた。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（NAKAHARA LAB: http://www.itn.brain.riken.jp 参照）。このような融合研究に挑戦する人々が増え、それが価値意思決定や報酬予測の研究だけでなく、多様な分野の研究で大きな発展につながることを期待している。また、本稿ではほとんど触れなかったが、意思決定の研究は社会的知性の研究と今後より関係を深めると思われる。それは、社会的意思決定、他人の心の予測、あるいは精神疾患のよりよい理解につながることだろう。神経科学が広範な科学と連携し、神経経済学や計算論的社会脳科学、計算論的精神医学、ひいては人間総合科学へと発展していくという、大きな学問的潮流が動きつつある。私たちもこの流れに貢献していきたいと考えている。

Temporal separator

図

図 1

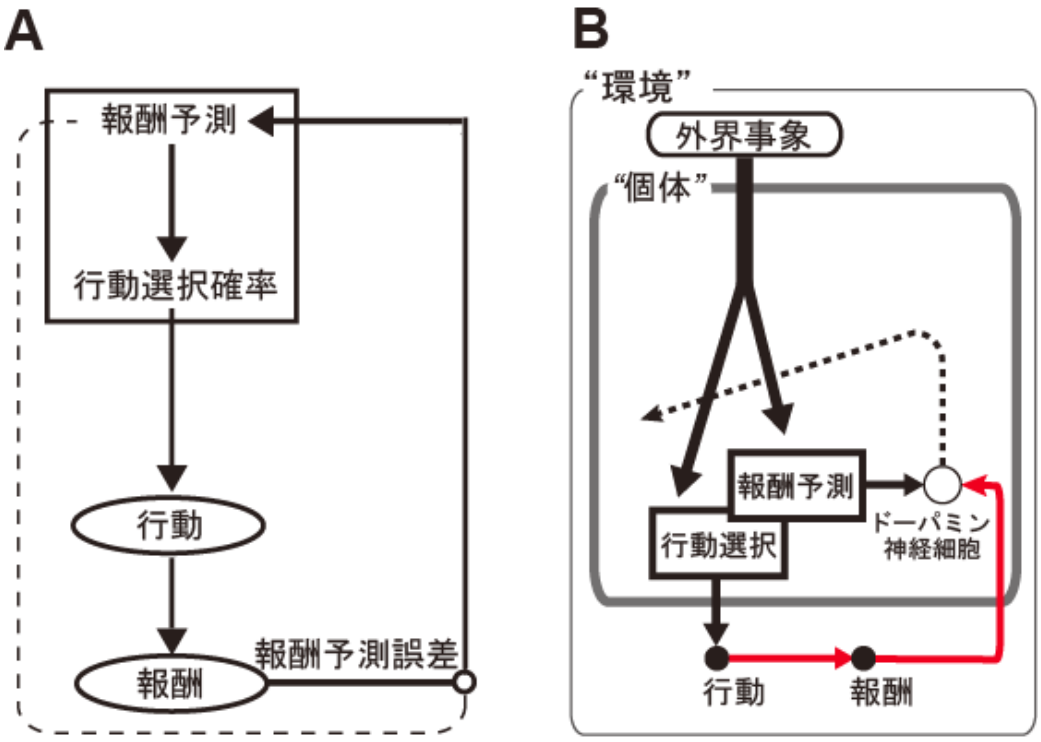

図 1.　モデルフリー強化学習(A)と報酬予測誤差仮説(B)

A.　入力からの直接連想により想起された報酬予測(価値)に基づいて確率的に行動を選択する。その結果として得られる報酬をもとに計算された報酬予測誤差(実際に得られた報酬量と予測していた量の差)を利用して、報酬予測を学習(更新)する。なお、四角の枠は個体の内部プロセスを表す。

B.　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)。この学習信号を利用して、太い黒色実線矢印の強度が変化することで報酬予測と行動選択の学習が行われる。

図 2

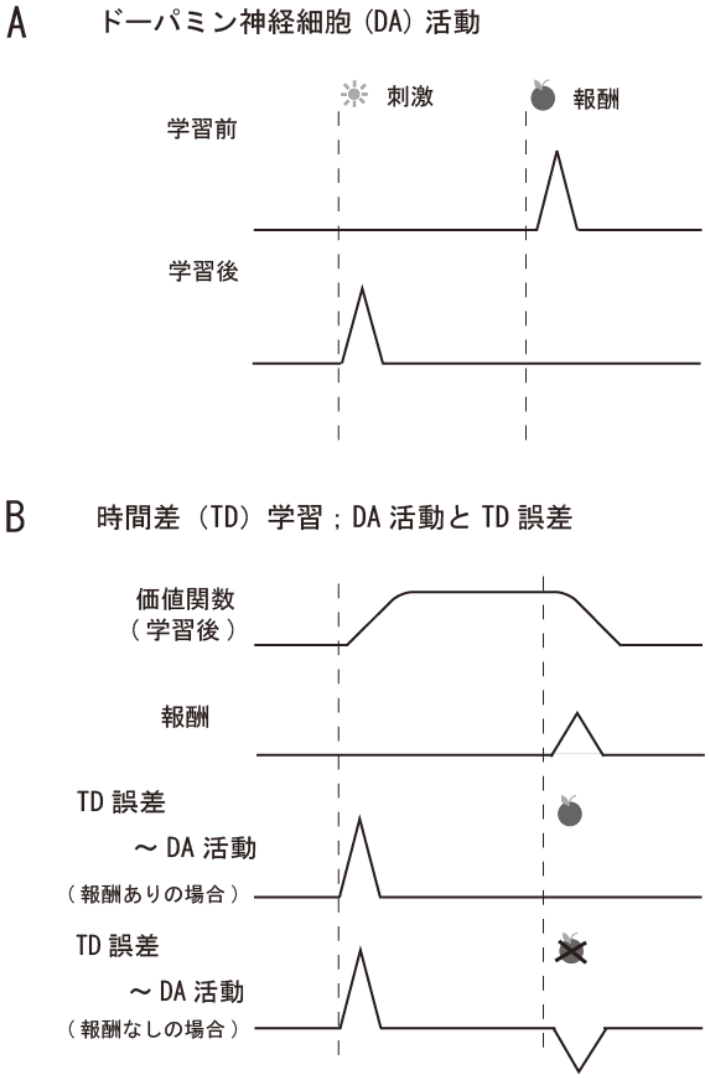

図 2. 刺激―報酬課題（古典的条件付課題）における、A;ドーパミン神経細胞活動(DA 活動)と、B;それに対応する時間差学習(TD 学習)および誤差(TD 誤差):B の下の 2 行の図は、ドーパミン神経細胞活動と TD 誤差が一致することを、報酬がもらえる場合ともらえない場合の両方について示している。

図 3

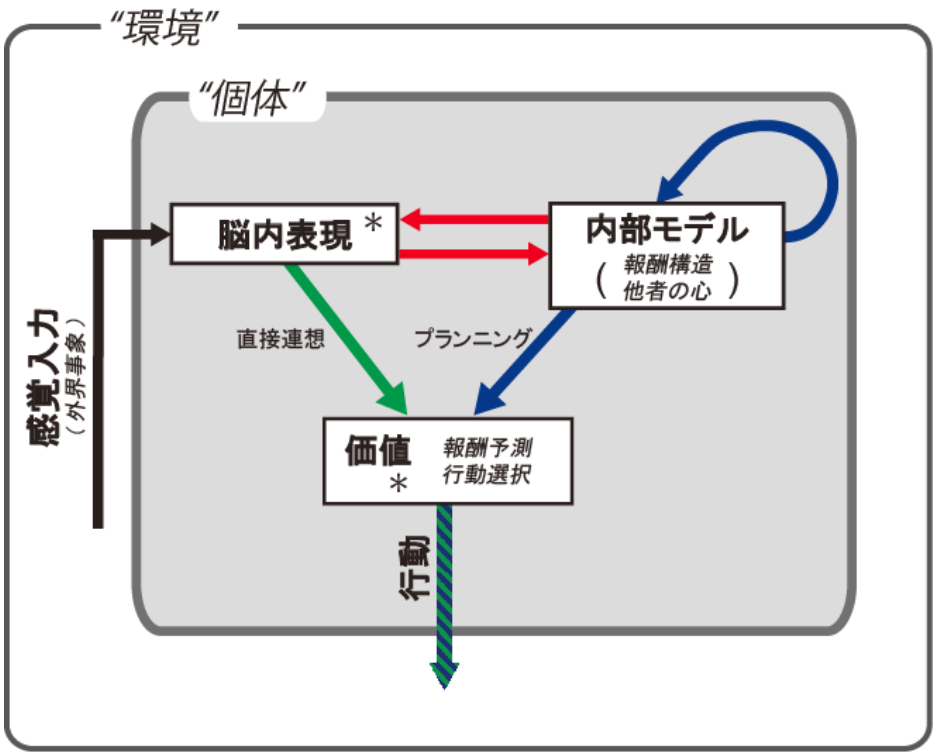

図 3. 価値に基づく意思決定のプロセス。エージェント、すなわち意思決定者が感覚入力（外界事象）を受けてそれが脳内表現になる。モデルフリーの強化学習はこの脳内表現からの直接連想によって価値を計算する（緑色矢印）。ドーパミン報酬予測誤差仮説では、さらに、この「感覚入力」が「脳内表現」に等しいと仮定したモデルフリー強化学習のモデルを用いる。一方で、経験を通じて、エージェントは報酬構造や他人の心などの内部モデルを構築する。この内部モデルを用いる意思決定では、プランニングや、索餌行動をはじめとするモデルベース強化学習などが可能となる（青色矢印）。内部モデルは外界を脳内シミュレーションする（再帰的な青色矢印）ことでこれらを実現する。赤色矢印は本文で述べた、脳内表現は内部モデルからの情報と感覚入力の統合であり、それによりモデルフリー強化学習がより強力になりうるという点を強調している。

図 4

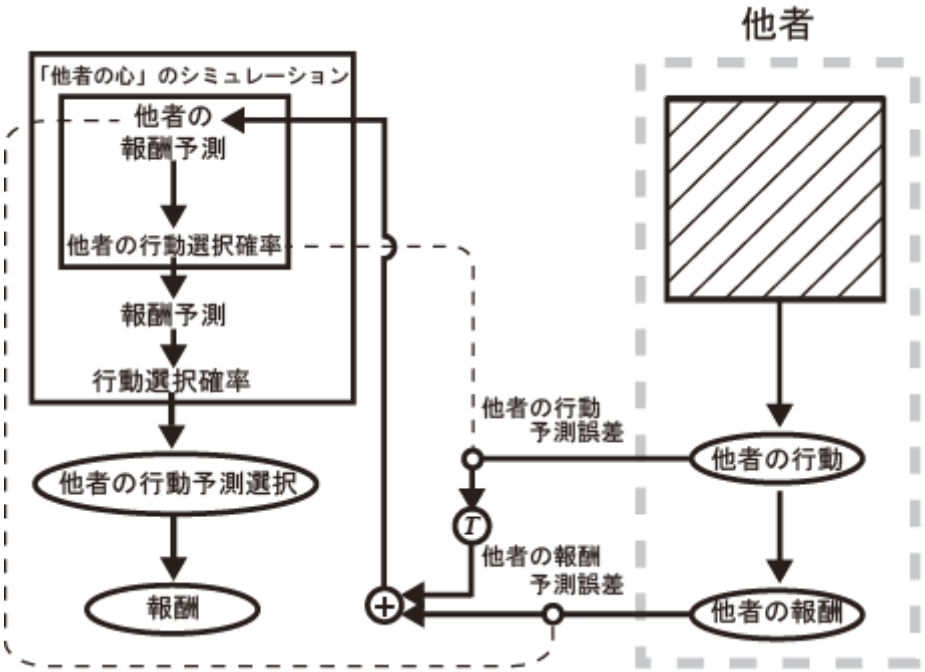

図 4.他者の心の脳計算モデルの概要

シミュレーション説と行動パターン説のハイブリッドモデル： 被験者は脳内で「他者の意思決定プロセス（報酬予測、行動選択確率）」を再現し、他者の行動を予測する（左図）。他者の実際の行動とその結果得られた報酬（右図）が明らかになった時点で、他者報酬予測誤差（他者が実際に得た報酬量と、被験者の脳内でシミュレーションしている他者が予測していた量の差）と他者行動予測誤差（他者の実際の行動とその予測の差）から「他者の報酬予測」を学習する。ここで、他者行動予測誤差は行動に関する予測誤差なので、そのままの形では報酬予測の更新には使えない。数学的には変分法と呼ばれる方法に対応した、行動予測誤差を報酬予測の学習に使える形に変換するプロセスが介在している。なお、斜線の四角枠は被験者から観察不可能な「他者の内部プロセス」を表す（図 1A を他者、あるいは本人が報酬予測を行って意思決定するプロセスと考えて参照するとよい）。

図 5

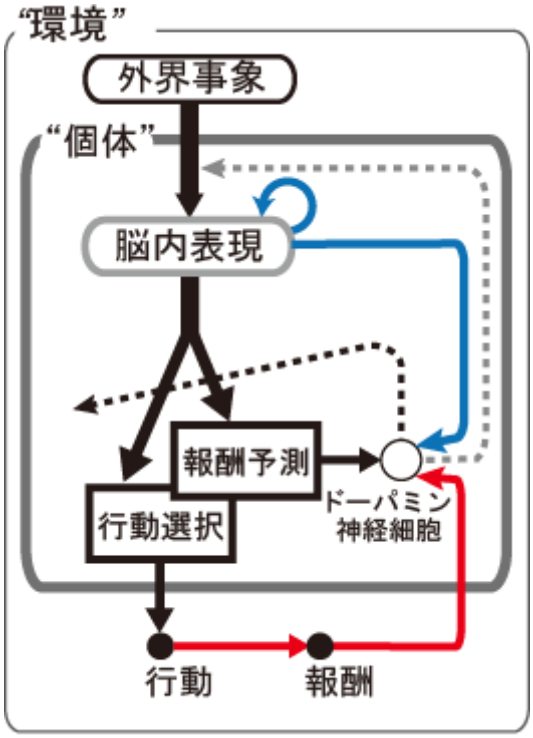

図 5. 報酬構造学習仮説：ドーパミン神経細胞活動は報酬構造を学習するための信号を表す（黒色点線矢印および灰色点線矢印）。ドーパミンや神経細胞は報酬予測のみならず、学習された報酬構造に関する入力（青色矢印）を受ける。さらに、ここでの報酬予測は各時点での外界事象と学習された報酬構造の両者を反映した脳内入力から生成される。この予測は原理的に報酬予測誤差仮説（図 1B）で用いられた報酬予測より優れている。この予測を利用した報酬予測誤差信号は、より優れた学習信号として報酬構造学習の信号の一部となる（黒色点線矢印）。報酬構造成分をより多く含むドーパミン神経細胞活動は、報酬構造を反映した内的表現の学習にも用いられる（灰色点線矢印）。

参考文献

References

1) Rangel A, Camerer C, Montague PR. A framework for studying the neurobiology of value-based decision making. Nature Reviews Neuroscience. 2008; 9: 545-556.

2) Glimcher PW, Rustichini A. Neuroeconomics: the consilience of brain and decision. Science. 2004; 306: 447-452.

3) Sutton RA, Barto AG: *Reinforcement Learning : An Introduction*. Cambridge, MA,: MIT Press; 1998.

4) Schultz W, Dayan P, Montague PR. A neural substrate of prediction and reward. Science. 1997; 275: 1593-1599.

5) 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, 深井朋樹, editors. 東京大学出版会; 2009. p. 159-221. 脳科学, vol 1.]

6) 中原裕之. 快楽が脳を創る. In: 脳研究の最前線. 脳科学総合研究センター 理, editors. 講談社; 2007. p. 233-297. vol 下.]

7) Hikosaka O, Bromberg-Martin E, Hong S, et al. New insights on the subcortical representation of reward. Curr Opin Neurobiol. 2008; 18: 203-208.

8) Doll BB, Simon DA, Daw ND. The ubiquity of model-based reinforcement learning. Curr Opin Neurobiol. 2012; 22: 1-7.

9) Rushworth MF, Noonan MP, Boorman ED, et al. Frontal cortex and reward-guided learning and decision-making. Neuron. 2011; 70: 1054-1069.

10) Kable JW, Glimcher PW. The neurobiology of decision: consensus and controversy. Neuron. 2009; 63: 733-745.

11) McDannald MA, Takahashi YK, Lopatina N, et al. Model-based learning and the contribution of the orbitofrontal cortex to the model-free world. Eur J Neurosci. 2012; 35: 991-996.

12) Dayan P, Niv Y. Reinforcement learning: The Good, The Bad and The Ugly. Curr Opin Neurobiol. 2008; 18: 185-196.

13) Daw ND, Niv Y, Dayan P. Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioral control. Nat Neurosci. 2005; 8: 1704-1711.

14) Wunderlich K, Dayan P, Dolan RJ. Mapping value based planning and extensively trained choice in the human brain. Nat Neurosci. 2012: 1-19.

15) Gläscher J, Daw ND, Dayan P, et al. States versus Rewards: Dissociable Neural Prediction Error Signals Underlying Model-Based and Model-Free Reinforcement Learning. Neuron. 2010; 66: 585-595.

16) Gallagher HL, Frith CD. Functional imaging of 'theory of mind'. Trends in Cognitive Sciences. 2003; 7: 77-83.

17) Suzuki S, Harasawa N, Ueno K, et al. Learning to Simulate Others' Decisions. Neuron. 2012; 74: 1125-1137.

18) Mitchell JP. Inferences about mental states. Philosophical Transactions of the Royal Society B: Biological Sciences. 2009; 364: 1309-1316.

19) Saxe R. Against simulation: the argument from error. Trends in Cognitive Sciences. 2005; 9: 174-179.

20) Behrens TE, Hunt LT, Woolrich MW, et al. Associative learning of social value. Nature. 2008; 456: 245-249.

21) Hampton AN, Bossaerts P, O'Doherty JP. Neural correlates of mentalizing-related computations during strategic interactions in humans. PNAS. 2008; 105: 6741-6746.

22) Nakahara H, Itoh H, Kawagoe R, et al. Dopamine Neurons Can Represent Context-Dependent Prediction Error. Neuron. 2004; 41: 269–280.

23) Schultz W. Updating dopamine reward signals. Curr Opin Neurobiol. 2013; 23: 229–238.

24) Bromberg-Martin ES, Matsumoto M, Hikosaka O. Dopamine in motivational control: rewarding, aversive, and alerting. Neuron. 2010; 68: 815-834.

25) Seitz AR, Dinse HR. A common framework for perceptual learning. Curr Opin Neurobiol. 2007; 17: 148-153.

26) Nakahara H, Hikosaka O. Learning to represent reward structure: A key to adapting to complex environments. Neuroscience Research. 2012; 74: 177-183.

27) 中原裕之, 鈴木真介. 意思決定と脳理論－人間総合科学と計算論的精神医学への展開. BRAIN and NERVE. 2013; 65: 973-982.