意思決定の計算神経機構

鮫島　和行

本稿では，脳を情報処理機械に見立てて理解する，神経科学における計算論的アプローチを解説する．特に，不確実な環境においてわれわれヒトを含む生物がどのように学習・適応していくのかを数理的にモデル化する強化学習の枠組みと，強化学習によって説明される大脳皮質と大脳基底核の神経回路の機能的役割についての研究動向を紹介する．計算論的神経科学が，動物の神経生理学的研究やヒトの脳の機能マップで得られるデータに対して，どのような解釈を与えるのかだけでなく，近年の脳神経操作手法における因果的な実証実験における数理モデルの有効性と展望について議論する．

キーワード：意思決定，強化学習，神経生理学

1. 計算神経科学

計算神経科学 (Computational neuroscience) とは，主に二つの立場からなる．一つは脳の仕組みを真似ることによって情報処理機械に活用する立場，もう一つは，脳を一種の情報処理機械としてなぞらえて見ることによって神経における生命現象を理解する立場である．前者は，既存の機械の情報処理方法とは違うなにか新しい数理的方法を探るために脳をのぞき込み，ヒントを得ようとする工学的立場であり，後者は，情報処理機械として脳を眺める窓として数理を使って脳を理解しようとする理学的立場ともいえる [1]．最近，パターン認識性能などで話題になっている deep network は前者の立場の例であろう [2]．一方で後者の例では，脳を既存の情報処理機械のアルゴリズムやその実装神経回路として捉えることにより，実際の神経回路ではどのような表現があり，その表現を使ってどのような行動が現れるのかを仮説として予測し，実際にそのような表現があるのかどうかを生理学手法によって検証することで，脳を理解する．しかし筆者は二つの異なる立場があるとはいえ，根は同じであると考える．脳をヒントに新しい数理的方法を発明することは，脳の原理を解明することと同義であるし，脳を理解するための窓を見つけることは新しく物事を捉える道具を発明することと同義だと考えるからだ．しかし，前者は，ひとたび新しい方法を見つけたらその方法が実際の脳で行われているのかどうかを検証する必要はない．これまで，私は脳を数理モデルで表現するという方法だけでなく，モデルによる情報表現やその変化が，実際の脳で実現されるのはどのようなメカニズムによっているのかを解明するために生理学的実験手法を用いて研究を行ってきている．特に，不確実な環境において，われわれヒトを含む動物がどのような意思決定をするのかを，定量的・数理的に定義し，ある評価基準を最適化する原理によって行動決定が行われると考え，その評価基準は何か？　それを達成するためのアルゴリズムや，実現している神経回路はどのようなものかを研究してきた．本稿では，このような意思決定の神経回路について，その計算メカニズムの理解の現状を解説する．

さめじま　かずゆき
玉川大学脳科学研究所・大学院脳科学研究科
〒 194–8610 東京都町田市玉川学園 6–1–1
samejima@lab.tamagawa.ac.jp

2. 不確実な環境における意思決定のモデル

「一寸先は闇」．日常は，次におきることを完全には予測できないという意味において，私たちは不確実な環境の中で生きている．そうはいっても，明日になれば一定の時間に朝は来るし，職場に行って日常を過ごすことになるだろう，という予測は立つ．「ある程度」予測ができるからこそ，過去の知識を利用し，過去にした失敗を繰り返すようなことはなく，「うまく」過ごすことができる．

この「ある程度」予測ができること，裏返せば「ある程度」予測が裏切られることや，「うまく」過ごすことを，定量的に表すために，環境の状態の不確実さの程度を確率変数として表現し，未来の状態を確率分布として表現する．また，「うまく」過ごすことを定量的に表現するために「報酬」を定義し，報酬ができるだけ得られるような過ごし方をすることが「うまく」いくことだと定義する．われわれ生物は，不確実な自然

環境の中での生存戦略を進化させてきたと考えることができる．一方，生まれてから，環境の知識を得て，社会の中で生き残る方法を学習によって獲得しているとも考えることができる．このように，数理的な定義に基づいて適応や学習の問題を考えることは，不確実な環境下における適応とその脳のメカニズムを考えるうえで重要である．一般的に神経科学は，外部からの観測に基づいて脳の中でなにがおきているのかを理解しようとする．一方で，計算論的神経科学では，脳がなにを目的に，どのようなアルゴリズムを用いているのかという仮説に基づいて，外部からの観測を理解しようとするアプローチであるともいえる．意思決定の問題では，生物が生存に必要とする食物や飲料，捕食者に出会う（出会わない）確率などを報酬や罰であるとして考え，その期待値を最大化するための状態から行動への変換則（方策）を経験のみから学習する問題であると，数理的に定義することによって，最適制御理論等の知見やアルゴリズムを援用して，脳の情報処理を理解することに相当する．

状態 s と行動 a と報酬 r という三つの変数を通して環境と相互作用し，長時間にわたる報酬の積算を最大化する問題設定は，強化学習と呼ばれている [3]．この問題設定は，行動出力そのものの修正を直接例示される教師あり学習とは異なり評価値のみから修正する必要がある．また，自ら探索的な行動をおこし「強化」されることによって修正されることからこう呼ばれている．最適制御との違いは，環境との相互作用を経験を通して学習するところにある．最適制御の場合，未来の状態の分布（状態遷移）や，報酬の与え方（報酬関数）は既知であるものとして，最適な行動を解くことによって「設計」することができるが，強化学習の場合は，状態の遷移や報酬の与えられ方そのものも経験から学習する必要がある．

この問題を解く方法（アルゴリズム）は，これまで多数提案されてきているが，二つに大別される．状態遷移や報酬関数を明示的に学習し，学習された内部モデルに基づいて行動を最適化するモデルベース強化学習と，明示的には内部モデルの形では学習せずに，それぞれの状態やそのときの行動から将来得られる報酬の期待値（価値関数）を学習する方法や，直接方策を学習する方法などのモデルフリー強化学習である [4]．モデルベース強化学習では，状態遷移が変化せずに報酬関数のみが変動した場合に，即座に行動を変化させることができる柔軟性があるが，行動するためには最適化計算を必要とするために毎回の行動に対する時間的・計算的コストが問題になる．一方で，モデルフリー強化学習では，直接方策または価値関数を学習し行動を行うために，行動への反応を素早く実行できるが，経験からの学習に多数の経験が必要となるため，環境変動に対する柔軟性は低くなるという特徴がある．

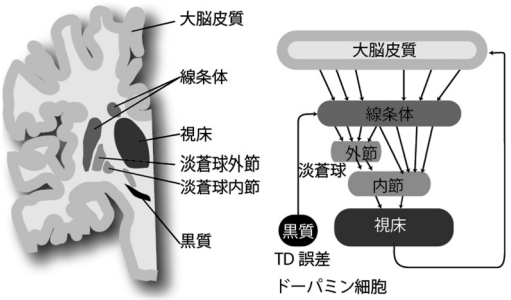

図 1　大脳基底核の神経回路の概略図．左は脳の前額断面のうち左脳だけを模式的に表している．右は複数ある大脳基底核の神経核と大脳皮質が構成するループ回路の模式図である．大脳皮質からの入力は線条体に送られ，線条体は淡蒼球内節へ直接投射する直接経路と，一旦淡蒼球外節を経て内節へとつながる間接経路とを通じて視床に投射する．視床は大脳皮質との間で相互につながりあっている．

3.　強化学習アルゴリズムと大脳基底核の価値表現

大脳基底核は，大脳の深部にあり複数の神経核（神経の塊）から成る（図 1）．大脳基底核は大脳皮質からの入力を受けていくつかの神経核を通して，脳幹から直接眼球や手足を制御する出力と，視床を介して再び大脳皮質の運動関連領野に戻るループ神経回路がある．大脳基底核の疾患の一つであるパーキンソン病では，動作の異常や最終的には無動などが発症することから，大脳基底核は運動制御系の一つであると考えられてきた．しかし，大脳皮質からの入力部である線条体という神経核は，大脳皮質のほぼ全領域からの入力のある神経核であり，運動ばかりではなく多様な認知的機能に関わることも報告されている．特に，学習に関しては，パーキンソン病患者において確率的な事象の予測問題に異常があることが報告されている [5]．このことからも，注意や作業記憶などの認知を含む行動やその学習に大脳基底核が関わるのではないかと考えられている．

線条体は，大脳皮質の全域からの入力を受ける領域であるが，同時に，中脳のドーパミン細胞からの入力もあり，脳内で最もドーパミン濃度の高い領域の一つである．中脳ドーパミン細胞において，モデルフリー

強化学習アルゴリズムの一つである Temporal Difference (TD) 学習則で用いられる誤差信号によく似た活動が，報告されている [6]．TD 学習則とは，将来得られる報酬の期待値である価値関数を学習するためのアルゴリズムの一つであり，最適制御理論でもよく用いられる Dynamic Programming (DP) アルゴリズムを，online で行う手法に相当する．離散時間の問題の場合，行動によって次状態 $s(t+1)$ への遷移が行われた場合に，前状態における価値関数 $V(s(t))$ を次の誤差信号によって更新するというものである．

$$\delta = r(t) + \gamma V(s(t+1)) - V(s(t)) \qquad (1)$$

ここで，γ は割引率であり，価値関数 $V(s)$ は将来の報酬を時間に応じて指数関数的に割いた総和の期待値である．TD 誤差は，割り引かれた次状態と前状態の価値の差分と現在得られる報酬 $r(t)$ の和で構成されている．すなわち，学習する前には報酬によって正の誤差が発生するが，価値関数が学習されれば，価値関数の差分によって相殺され誤差は発生しなくなる．その代わり，報酬を予期できる状態変化によって誤差が発生するように時間的に前へと誤差が移動することになる．また，もし価値関数によって予期された報酬が得られなかった場合には，負の誤差が生じることになる．中脳のドーパミン細胞では，まさにこれと同じ信号が観測されている [6]．

ドーパミン細胞は，主に線条体と大脳皮質の前頭葉へとその出力を送っている．パーキンソン病ではドーパミン細胞が死滅し，線条体のドーパミン濃度が下がる．パーキンソン病の治療薬として開発された人工ドーパミンである L-DOPA の投与の有無によって，未知の刺激と報酬の間の連合学習に影響があることや [7]，健常者においても，L-DOPA の投与とドーパミン受容体のブロッカーであるハロペリドールの投与で同様に視覚的に与えられる記号と報酬との連合学習に影響があることが報告されている [8]．

また，このような刺激と報酬との連合学習中に TD 誤差と fMRI による線条体の脳活動が相関していることも報告されている．脳活動は物理的には血流を計測するため，ドーパミンによる神経活動そのものを計測しているわけではない．また，時間的にも，刺激のタイミングと報酬のタイミングが近ければ，どちらのタイミングで神経活動がおきているのかを分離することは困難になる．

筆者はこれまでに，動物実験によって線条体からの単一細胞の神経活動記録を，強化学習を行っている際に

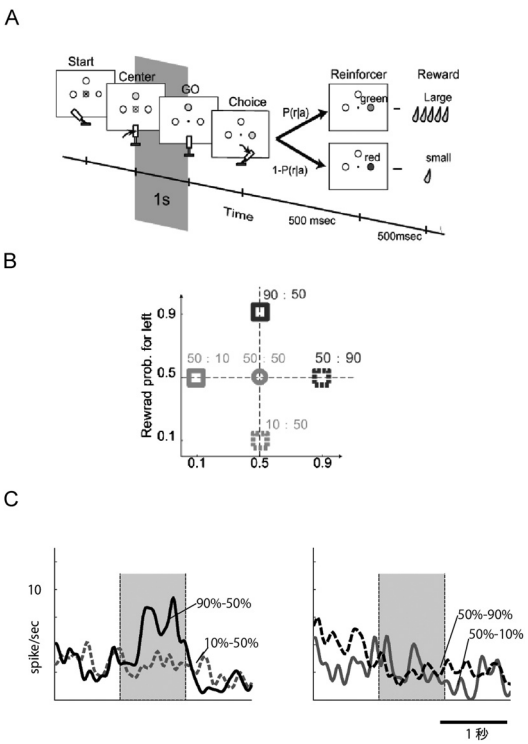

図 2　動物に行わせた意思決定課題と，神経活動の一例 [9] より一部改変．A 動物に行わせた課題．B 課題に用いた報酬確率の組み合わせ．C 神経活動の例

行ってきた．このような生理学的実験によって，線条体の神経細胞が価値関数と似た信号を意思決定の直前にしていることを報告した [9]．線条体の活動は，様々な刺激・運動・報酬のタイミングで活動がみられるが，この実験では，運動の直前に意思決定を行っていると思われる情報の入力や運動を行っていない期間に着目して神経活動を解析した．動物は握っているレバーを左右に倒すことで確率的に報酬を得る課題を遂行する（図 2A）．このとき，報酬が得られる確率を 3 種類（10%，50%，90%）設定し，一定期間ごとに変化させ，動物に学習させた（図 2B）．また，左右の報酬確率は差が 40%異なっており，常にどちらかが有利な条件となっている．このとき，ある線条体の単一神経細胞活動は，図 2C のような活動を示した．すなわち，仮に左を選択したときの報酬の確率が 90%のときに 10%である場合よりもより強く活動したことになる．一方で仮に右を選択したときの報酬が変動しても神経活動は変化していないことになる．このことから，神経細胞活動が，ある行動 $a(t)$ を仮に選んだとしたときの行動価値

$$Q\left(s(t),a(t)\right)=E[\sum_{k=0}^{\infty}\gamma^{k}r(t+k)|s(t),a(t)]$$

と似た活動を示していることになる．この行動価値は次の TD 誤差によって学習することができる．

$$\delta=r\left(t\right)+\max_{a\in A}Q(s\left(t+1\right),a)-Q(s\left(t\right),a\left(t\right))$$

ここで，A は取りうる行動の集合，$s(t)$，$s(t+1)$，$a(t)$ はそれぞれ前状態，および次の状態，時刻 t でとった行動を表す．

線条体に入力しているシナプスの可塑性が，ドーパミンの濃度に依存して変化しているという報告がされている [10]．すなわち，ドーパミンに依存して，線条体の細胞の入力から出力までが変化し，刺激から報酬を予期することを学習し意思決定に寄与するのではないかという情報処理モデルを考えることができる．すなわち，TD 誤差によって行動価値関数が線条体において学習される，というモデルである．これまでに，手の運動ばかりではなくサルの眼球運動 [11] や，ラットのノーズポークによる意思決定課題での単一神経活動と価値関数との相関が報告されている [12]．また，ヒトの fMRI を用いた実験でも同様に価値関数や価値関数に基づく誤差信号と相関した脳活動が線条体において報告されており，価値に基づく意思決定が大脳皮質と大脳基底核を含むループ神経回路で実現されている可能性が示唆されている．これらの報告では，なんらかの強化学習モデルを仮定し，被験者や被験体の意思決定の結果である行動選択をもっとも説明するモデルパラメータを推定し，そのモデルパラメータにおいて進行する価値関数の学習や誤差信号との相関を求めている．いわば，脳をある強化学習アルゴリズムで学習する機械になぞらえて見たときに，一見外部から観測できない内部変数を仮定に基づいて推定し，その内部変数と脳の信号との相関を求めていることに相当する．このようなモデルに基づく神経活動の解析によって，直接外部で観測できる変数との一致を見ることができないような内部変数の解釈が可能になる．

4. 意思決定における神経回路操作による因果的証明

しかし，モデル変数と神経活動との相関による研究のみでは，因果的な証明を行うことは不可能である．因果的証明を行うためには，内部変数を表現している特定の神経発火表現を操作することによって，意思決定にどのような影響が現れるのかを検証する必要がある．

脳を操作することによる因果的証明は，これまでの神経科学でも多く試みられてきている．ヒトの脳を操作する研究は，外部から強力な磁界を発生させ脳に一過性の電流を発生させることによる領域の遮断や促進方法は TMS と呼ばれ多くの研究があるほか，頭表に電流を流し頭蓋内の皮質に微弱な電位変化を発生させることで，神経回路のなんらかの状態を変化させる研究などが近年注目されている．動物実験では，神経回路の破壊や冷却等の方法のほか，直接局所への注入などが行われている．また，薬理的操作は，シナプスや神経膜上に存在する受容体やトランスポーターの働きに変化を与えることで操作することができる．しかし，これらの手法はいずれも神経回路の大きな領域での操作であり，ミリメートル単位の細かな神経集団に対して，特定の神経回路を操作することに限界があった．

近年，光遺伝学的手法を用いて，特定の遺伝的に標識された神経細胞の活動電位を発生させたり，遮断したりすることが可能になった [13]．脳は一つの領域の中にも，多種類の神経細胞が混在しており，それぞれが異なった分子を発現し，それぞれが異なる領域に接続した複雑な神経回路を構成している．これまでは，たとえ非常に微小な領域に対して電流注入や薬物注入を行ったとしても，その周辺にいる全ての種類の神経細胞に影響を与えてしまうために，神経回路の特定の配線だけを操作することは困難であった．この技術を用いて，線条体の細胞を種類ごとに分けて操作を行うことで，意思決定への因果性が証明されている．

線条体では，大脳基底核の出力核である淡蒼球外節や黒質網様部へ接続する直接経路を形成する神経細胞と，淡蒼球内節を経由して出力核である淡蒼球外節へと接続している間接経路の形成する神経細胞が混在している．直接経路を形成する神経細胞は主にドーパミン D1-type 受容体が発現しており，間接経路の細胞は D2-type 受容体が発現していることが知られている．このことを利用し，直接経路だけに光感受性のチャネルを発現させることによって，レーザーによって直接経路と間接経路をそれぞれ独立に操作することができる．

線条体と淡蒼球の神経細胞は，抑制性の神経細胞である．したがって，直接経路は視床の神経細胞活動に対して抑制している出力をさらに抑制することで，脱抑制することが知られている．一方で，間接経路は線条体から淡蒼球外節–内節–視床とつながるため，3 重の抑制となる（図 1 右）．この直接経路と間接経路の神経細胞は出力系への影響が拮抗する系と考えることができる．もし，これらの細胞活動が価値を学習・表現す

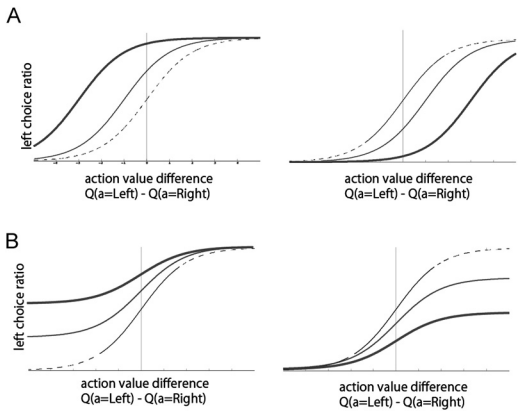

図 3 左半球と右半球の線条体の直接経路と間接経路をそれぞれ光遺伝学的手法によって刺激しわけた結果 [14] の模式図．A 左パネル：右半球の直接経路および左半球の間接経路を刺激した場合のマウスの左選択確率，A 右パネル：左半球の直接経路および右半球の間接経路を刺激した場合の結果．B 行動選択確率にバイアスがのると仮定した場合の予測を表す．点線は刺激しない場合の選択確率の変化，線の太さが太いほど刺激強度が高い．

るのであれば，直接経路は特定の行動価値を上げてその行動選択を促進することに，間接経路は行動価値を下げて行動選択を抑制することに寄与するはずである．また，線条体は，対側への行動を実行するときにより活動するという対側優位性が知られている．もし，左の線条体の直接経路の神経細胞が $Q(s, a{=}\ \text{right})$ を，間接経路が $-Q(s, a{=}\ \text{right})$ と正負の行動価値を表現し，反対に右の直接経路が $Q(s, a{=}\ \text{left})$，間接経路が $-Q(s, a{=}\ \text{left})$ が多く存在するのであれば，左右の線条体に光ファイバーをインプラントして左右独立に直接経路と間接経路を制御することによって，行動選択を制御できるはずである．

Tai らは，この実験をマウスの線条体において行い，左線条体の直接経路と右線条体の間接経路を刺激することによって右の行動選択が増加し，逆に左線条体の間接経路と右線条体の直接経路を刺激することによって左の行動選択が増加することを報告している [14]．図 3 に示すように，左右の行動価値にガウスノイズが重畳してその競合によって行動選択がおきるモデルを考えると，競合する価値の差を横軸にとった選択確率のカーブは，S 字型のシグモイド関数になる．これをボルツマン選択と呼ぶ．光刺激によって，右線条体の直接経路と左線条体の間接経路を刺激した場合，選択確率のカーブは負の方向にシフトした（図 3A 左）．これは $Q(s, a{=}\ \text{left}) - Q(s, a{=}\ \text{right})$ は擬似的に増加したと解釈することができる．また右線条体の間接経路と右線条体の直接経路を刺激した場合は，逆に正方向にシフトした（図 3A 右）．これは $Q(s, a{=}\ \text{left}) - Q(s, a{=}\ \text{right})$ を擬似的に減弱させたと解釈することができる．

この実験で重要なことは，刺激をすると常に同じ割合で一定方向の行動がおきるというモデルでは説明できないことにある．もしそうであるならば，刺激の強度が増すにつれて，上方向または下方向へのシフトが見られるはずであると理論的には予測できる（図 3B）．しかしこれは少なくともマウスの脳ではおこらなかった．これは，線条体ではなくほかの領域ですでに意思決定が行われ，その結果が送られて実行する系に対する操作が加わったのではなく，意思決定をする前段階の価値の情報に対して操作が行われ，それが意思決定に反映されていることを示唆している．

5. 最後に

意思決定の計算神経機構について，特にモデルフリー強化学習のアルゴリズムが，大脳基底核の情報表現や強化学習の神経によるメカニズムの理解に有効であることを，いくつかの研究を題材に紹介した．このような，脳が対処している問題を定式化し，数理的な解析を行うことで一定の解を求め，その解を実現するいくつかの方法が，現実の脳でおきている現象の理解に貢献していることを示した．

しかし，現存のアルゴリズムでは説明できない現象が多数存在することや，脳は単一のアルゴリズムのみで意思決定を実現しているわけではなく，状況に応じていくつかのアルゴリズムを使い分けることも考えられる．実験室の環境では，同じタスクを実行し確率的に与えられる液体報酬や個体報酬の摂取を繰り返す中で，効率よく行動することを要求するようなクリーンな状況を想定している．自然環境の中では，柔軟性を要求されるような環境の素早い変化や，一部のみの変化から一般化するための能力が要求されたり，場合によっては，知識そのものが役には立たないが，探索戦略のみが役に立ったりする場合も多い．モデルフリー強化学習アルゴリズムが最適解を導けるような課題ばかりではない．近年では，外界のモデルの学習や，特定の事象（エピソード）から推論をする能力と，意思決定の関係性が研究され，モデルベース強化学習がどのように脳内で実現されているのかについて，理解が進みつつある．

現状では，意思決定の一つのモデルとして強化学習という枠組みとそのアルゴリズムが有用ではあるが，

それは真実である保証はない．ほかの科学分野と同じように，モデルとは矛盾する実験結果の発見とそれらを説明するモデルの更新を繰り返すことによって定量的に理解を進めていくことが重要ではないだろうか．

参考文献

[1] D. Marr, *Vision*, MIT Press, 1982 (reprinted in 2010).

[2] G. E. Hinton, S. Osindero and Y. W. Teh, "A fast learning algorithm for deep belief nets," *Neural Computation*, **18**, pp. 1527–1554, 2006.

[3] R. S. Sutton and A. G. Barto, *Reinforcement Learning An Introduction*, MIT Press, 1998.

[4] N. Daw, Y. Niv and P. Dayan, "Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioral control," *Nature Neuroscience*, **8**, pp. 1704–1711, 2005.

[5] B. J. Knowlton, J. A. Mangels and L. R. Squire, "A neostriatal habit learning system in humans," *Science*, **273**, pp. 1399–1402, 1996.

[6] W. Schultz, P. Dayan and P. R. Montague, "A neural substrate of prediction and reward," *Science*, **275**, pp. 1593–1599, 1997.

[7] M. J. Frank, L. C. Seeberger and R. C. O'reilly, "By carrot or by stick: Cognitive reinforcement learning in parkinsonism," *Science*, **306**, pp. 1940–1943, 2004.

[8] M. Pessiglione, B. Seymour, G. Flandin, R. J. Dolan and C. D. Frith, "Dopamine-dependent prediction errors underpin reward-seeking behaviour in humans," *Nature*, **442**, pp. 1042–1045, 2006.

[9] K. Samejima, Y. Ueda, K. Doya and M. Kimura, "Representation of action-specific reward values in the striatum," *Science*, **310**, pp. 1337–1340, 2005.

[10] J. N. J. Reynolds and J. R. Wickens, "Dopamine-dependent plasticity of corticostriatal synapses," *Neural Networks*, **15**, pp. 507–521, 2002.

[11] B. Lau and P. W. Glimcher, "Value representations in the primate striatum during matching behavior," *Neuron*, **58**, pp. 451–463, 2008.

[12] M. Ito and K. Doya, "Validation of decision-making models and analysis of decision variables in the rat basal ganglia," *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience*, **29**, pp. 9861–9874, 2009.

[13] K. Deisseroth, "Optogenetics," *Nature Methods*, **8**, pp. 26–29, 2011.

[14] L. Tai, A. Lee and N. Benavidez, "Transient stimulation of distinct subpopulations of striatal neurons mimics changes in action value," *Nature Neuroscience*, **15**, pp. 1281–1289, 2012.

