別紙 4

報告番号	※ 第 号

主　論　文　の　要　旨

論文題目 Triple processes underlying human decision making in reversal learning tasks: Functional significance and evidence from the computational model fit to human behavior
(逆転学習課題における意思決定の 3 つの処理過程－その機能的意義および証拠についての計算論モデルを用いた行動データ適合による検討－)

氏　　名 白　宇　(BAI　Yu)

論　文　内　容　の　要　旨

本論文は、変化する環境における人間の適応的な学習過程のメカニズムについて、報酬予測誤差の更新、注意の制御および方略の調整という三つの過程のダイナミックな結合によって実現されることを強化学習計算論モデルと生理指標を用いて実験的に検討を行ったものである。

第 1 章では、強化学習モデルの一般的原理およびその神経基盤を説明し、複雑な意思決定過程において強化学習モデルを応用した関連研究を紹介した上で、一般的な強化学習モデルの持つ二つ限界を指摘した。強化学習とは正解を教示されることなく、エージェントは自分が取った行動に対しする報酬および罰のような強化信号を手がかりにして試行錯誤で行動を最適化する学習過程である (Sutton and Barto 1998)。一般的な強化学習モデルでは、報酬予測誤差 (Reward Prediction Error: RPE) に基づいて行動価値を更新し、最適行動に到達する過程が想定される。RPE の更新規則は動物心理学で知られる Rescorla-Wagner モデルをベースにしたものがある: (例えば Q-learning モデルなど)。しかし、Q-learning モデルでは二つの限界があると考えられる。一つは連合可能性(associability)の更新の無視であり、もう一つは変動可能性(variability) の更新の無視である。連合可能性は Q-learning モデルの中において学習率で表現されている。変動可能性は逆温度で表現されている。この二つのパラメータは両方とも定数として設定され、更新が想定されていない。しかし、いくつかの研究によって、この二つのパラメータが学習過程において変動可能であることを示している (Doya 2002; Li et al. 2011; Pearce and Hall 1980)。その上に、それらのパラメータに関連する生理基盤も明らかになりつつある (Aston-Jones et al. 1994; Roesch et al. 2012)。本研究では上記のこの二つの限界を解決できるような新しいモデルを提案することを目的とする。

最近の研究では、注意機能の変化プロセスを RW モデルに取り入れた hybrid モデルが提唱された（Li et al. 2011)。hybrid モデルが RPE の更新と注意の変化を表す学習率の更新の二つのプロセスから構成される。この hybrid モデルが古典的条件づけ課題において、Q-learning モデルより優れていることが示されているが（Li et al. 2011)、その機能的な意義についてまだ明確になっていない。また hybrid モデルで仮定された二つのプロセスの中の RPE の更新に関する処理は中脳ドーパミン信号によって表現されることが多くの研究によって示されてきた。学習率の更新について扁桃体の活動が関与していることもいくつかの研究によって示されてきた（Roesch et al. 2012)。一方で、この二つのプロセスに関する神経基盤に関する重要な電気生理学的な知見も存在する。例えば、Feedback-related negativity（FRN）という事象関連電位（ERP）の成分は、負の結果を示す外部刺激によって惹起される。この FRN および FRN に後続する陽性 ERP 成分の P300 が RPE と関連することがエラー関連電位－強化学習理論（ERN-RL）が提唱されている（Holroyd and Coles 2002)。たが、これらの ERP 成分 が RPE との関連することが多くの研究によって示されている一方で、RPE との間に関連性が検出されていないという報告も数多く存在する（Hajcak et al. 2007)。さらに、最近いくつかの研究ではこれらの ERP 成分が RPE だけではなく、学習率の更新にも寄与している可能性を示している（Behrens et al. 2007)。

そこで、本論文ではまず先行研究によって提唱された hybrid モデルが一般的な RL モデルである Q-learning モデルと比較して,行動選択課題においてどのような機能的な優位性（報酬随伴性の高い刺激をより多く選択できること）があるのかを検討する。また、hybrid モデルで仮定された二つのプロセスと関連する可能性のある神経科学の既存知見（エラー関連 ERP 成分）との関連性を検討し、hybrid モデルの生理学的妥当性を確認する。最後に hybrid モデルを基に、さらに逆温度更新可能な新しいモデルを提案した。これらの目的を順次に第 2 章、第 3 章と第 4 章で検討する。

第 2 章では、hybrid モデルの機能的意義について、コンピュータシミュレーションおよび行動データのモデルベース解析を用いて検討を行った。実験課題として、ダイナミックな適応行動を研究する際によく用いられる逆転確率学習課題（Probabilistic reversal learning task）を使用した。この課題では、エージェントに二つの選択刺激を提示し、その中から一つを選ぶように指示する。二つ刺激は高い報酬随伴率を有する有利刺激と低い報酬随伴率を有する不利刺激によって構成される。この随伴関係が課題の途中で逆転する。エージェントはこれらの刺激に関する情報を一切与えられず、ただ試行錯誤の中で自分にとって有利な行動を取ることだけが明示される。このような課題をよりよく遂行できるには、エージェントにいくつかのことが要求される。第一に有利刺激と不利刺激をすばやく弁別するこ

と；第二に弁別ができたあとに有利刺激を多く選択すること；第三に有利刺激を選びつつも逆転が生じるときに素早く反応を切り替えることである。仮に hybrid モデルで仮定した注意調整プロセスの関与が機能するならば、hybrid モデルに基づいた行動はより有利刺激が選択される結果をもたらすはずである。コンピュータシミュレーションの結果では、hybrid モデルがよりよい課題成績が得られことを示し、

特に極端な初期状態で学習を開始する場合、Q-learning システムでは有利刺激を選ぶ確率は hybrid モデルよりも低いことが明らかになった。したがって、研究 1 では、hybrid モデルは学習率を調整することによって、Q-learning モデルよりも逆転学習のような複雑な状態において強い適応性があることがコンピュータシミュレーションによって示された。これに加え、シミュレーションにおいては難易度の異なる 3 つの条件が設定され、より難易度の高い条件ではより高い行動適応性が求められる。ほかの 2 条件と比べて hybrid モデルは最も難易度の高い条件において Q-learning モデルよりよい選択を選択した回数が多かったことはシミュレーションの結果によって示された。さらに、ヒトを参加者に同課題を使用した実験で得られた行動データに hybrid モデルと Q-learning モデルでパラメータフィットを行った結果、hybrid モデルは Q-learning モデルよりも行動データをよりよく説明できることが明らかになり、hybrid モデルで想定されたプロセスの存在が実際の行動データ解析によって確認できた。

第 3 章では、ERP 成分が hybrid モデルで仮定された RPE と学習率の更新にどのように関与するかを調べるため、実験で得られた行動データを hybrid モデルに基づいてパラメータ推定を行い、各試行の RPE および学習率の更新差分を算出した。その算出した RPE および学習率を基に ERP 成分との関連を調べた。FRN の振幅と負の RPE の値に有意の相関が見られ、P300 の振幅は正の RPE の値と強い相関が見られた。これらの ERP 成分は学習率との関連は見られなかった。これらの結果から、FRN と P300 が異なる報酬価における RPE を反映することが推測され、hybrid モデルで提唱した RPE 関連プロセスのみを反映すると考えられる。

第 4 章では、hybrid モデルおよび Q-learning モデルでは選択行動の柔軟性に関するパラメータがあり、それは逆温度（inverse temperature）というパラメータである。逆温度は、あらゆる可能性をモニタするために多くの可能な選択肢を均等的に選ぶという探索行動(exploration) と、単一の良い選択し固執し、効用の最大化を測る搾取行動(exploitation)のバランスを表現すると言われている（Doya 2002; Ishii, Yoshida, and Yoshimoto 2002)。hybrid モデルの中では逆温度の値が定数として設定されており、課題中に状況が変化するにもかかわらずエージェントが探索行動と搾取行動のバランスを調整しないことになっている。したがって、研究 4 では第 2 章と第 3 章で検討した hybrid モデルをベースに、逆温度の更新プロセスを想定したモデルを提案し、その新しいモデル（β-hybrid モデル）の機能的意

義について、コンピュータシミュレーションを用いて hybrid モデルと比較ながら検討することにした。連続逆転学習課題において、高次機能の一つである方略転移（strategy transfer）が起こることが知られている（Rygula et al. 2010)。逆転学習課題の 5 ブロックの最初のブロック（方略転移が生じていない試行）と最後のブロック（方略転移が生じた試行）で得られた行動データをそれぞれ β-hybrid モデル hybrid モデルでパラメートフィットさせることで、両モデルを比較した。シミュレーションの結果では、β-hybrid モデルは hybrid モデルよりも報酬随伴性の高い刺激の選択率が高いことが示された。また hybrid モデルでは初期値が極端な場合に学習が進まないという結果を得たが、β-hybrid モデルではたとえ初期値が極端な場合でも学習が進むことが確認された。また研究 1 と同様、シミュレーションでは難易度の異なる 3 つの条件が設定され、β-hybrid モデルが難易度の高い条件では顕著に hybrid モデルよりよい成績を示していた。また、行動データにおいては、方略転移が確認され、方略転移が生じたブロックのデータにおいて、β-hybrid モデルが高い説明力を持っていることが明らかになった。このように、β-hybrid モデルがより高次な機能を必要とする学習過程を反映できることが示された。

最後に第 5 章では，第 2 章から第 4 章までの実験的検討から得られた知見を総括した上で、hybrid モデルのプロセスを包含する β-hybrid モデルの機能的意義およびそれに関連する可能性のある神経基盤について議論を行った。これらの知見により、β-hybrid モデルで表現された三つのプロセス、RPE の更新プロセス、学習率の変化のプロセスと逆温度調整のプロセスは、複雑な環境における学習行動に大きく寄与すると考えられた。また、計算論モデルの式の通り、この三つのプロセスが独立な下位システムではなく、ダイナミックに結合された一つの全体として機能することが示された。一方で、本研究で行った議論はすべて単一の課題におけるものであり、その一般的な妥当性および適応性について、今後の研究で本研究と異なる課題を用いて検討する必要がある。また、類似のシステムを想定して、異なる表現法を用いた計算論モデルも存在するため、モデル間の異同についてもさらに検討を行うことが今後の課題として議論された。