論文内容の要旨

博士論文題目

Multiple Reinforcement Learning Action Selection Strategies in Prefrontal-Basal Ganglia and Cerebellar Networks

（強化学習理論に基づく意思決定戦略における前頭前野－大脳基底核－小脳系神経回路の計算論的機能に関する研究）

氏名

Alan de Souza Rodrigues

（論文内容の要旨）

Humans can learn to select actions from scratch, by trial-and-error, or by using knowledge from previous experiences. Also, a thoughtful deliberative strategy can be used for planning of future behaviors, whereas a reactive strategy for selection of learned sensory-motor mappings. Reinforcement Learning (RL), a computational theory of adaptive optimal control, proposes two formal methods for action selection: Model-Free (MF) uses action values to predict future rewards based on current states and available actions and Model-Based (MB) uses a forward model by mental simulation to predict future states reached by hypothetical actions. It has been suggested that humans might use MF and MB strategies which are implemented locally by distinct brain areas or in neural networks linking the prefrontal cortex, basal-ganglia and cerebellum. The goal of this thesis is to investigate whether and how humans use multiple RL action selection strategies, and where in the brain they are implemented.

We hypothesized based on RL a parallel model for action selection and learning from reward and trial-and-error. When knowledge of the task dynamics is unavailable, a MF valued-based method is used for exploration and learning predictions of future rewards from sensory states and actions. If an internal model exists, a MB forward method is implemented by mental simulation to predict future states reached by hypothetical actions. After repeated experiences, a MF memory-based method is employed for fast selection of actions found successful for a given sequence of sensory states.

This hypothesis was experimentally tested as human subjects performed a ‘grid-sailing task’ and learned to move a cursor from a start position to a target goal with the shortest finger movement sequence by button pressing. Performance feedback was provided as a reward score at the trial end. Subjects extensively learned action sequences for fixed key-mappings (KM) and start-goal (SG) sets in a training session and were later tested under three task conditions: a) Condition 1, new KM, b) Condition 2, learned KM, and c) Condition 3, learned KM-SG sets, practiced in the training session. The response start time was manipulated by introducing a delay period of about 4~6s preceding the response start go signal in half of the trials.

Behavior analysis revealed distinct performance profiles in the test session: a) exploratory, variable and slow learning in Condition 1 due to the use of the new KM; b) fast learning and high reward score especially in trials with the delay period in Condition 2 as subjects could use the learned KM, c) accurate performance with fast sequence execution in Condition 3 which required the learned KM-SG sets in the training session. These results suggest that a value-based exploratory method was used in Condition 1; a model-based planning method in Condition 2 and the memory-based habitual method in Condition 3.

The fMRI analysis of the brain signals during the delay period preceding movement execution revealed predominant activity by the use of: a) value-based method in Condition 1 in the medial orbital frontal cortex, ventromedial striatum and left cerebellum; b) model-based method in Condition 2 in the dorsolateral prefrontal cortex, dorsomedial striatum and right cerebellum, and c) memory-based method in Condition 3 in the supplementary motor area, right dorsolateral striatum and right anterior cerebellum.

We could demonstrate by the behavior analysis that humans use multiple action selection strategies depending on their experiences with a given task, existence of an accurate internal model and available time for thinking and use of the model. The fMRI results showed that an action selection method is implemented in the brain by interaction of multiple brain areas, such as networks linking the prefrontal cortex, the basal-ganglia and cerebellum, and not locally as had been previously speculated. What computational operations these brain regions play in mental simulation, planning and preparation of future behaviors and how they are disrupted in psychiatric conditions is still a problem to be elucidated.

氏　名	Alan de Souza Rodrigues

（論文審査結果の要旨）

ヒトの行動選択は，試行錯誤により学習する場合やこれまでの知識を利用する場合などがある．これらはいずれも強化学習の枠組みで説明することができ，それぞれモデルフリー強化学習，モデルベースド強化学習と呼ばれる．本研究は，この仮説をヒトを用いた行動実験によって実証するとともに，fMRI を用いた脳活動計測により，行動選択に関与している部位を特定したものである．

本研究では「グリッドセイリングタスク」を用いる．キーを押してカーソルをスタートからゴールまで移動させる課題であり，押したキーとカーソルの移動方向の関係（KM）は既知または未知である．また，スタートとゴールの位置（SG）も既知または未知である．

行動実験の解析により，未知 KM 未知 SG の場合(A)には，モデルフリー強化学習と符合する成績カーブが得られた．既知 KM 未知 SG の場合(B)には内部モデルによる予測が可能となるため，モデルベースド強化学習と符合する成績カーブが得られた．また，既知 KM 既知 SG の場合(C)には覚えている解を思い出すメモリベースの戦略が可能であるため，非常に早くよい成績を得るという成績カーブが得られた．

また，それぞれの条件における脳活動を fMRI で調べ，各試行の KM と SG の提示から行動開始までの準備期間中に(A)では内側眼窩前頭皮質，腹内側線条体および左小脳，(B)では背外側前頭前皮質，背内側線条体および右小脳，(C)では補足運動野，背外側線条体および右前小脳における賦活を明らかにした．

以上をまとめると，本論文は，ヒトの行動選択の 3 種類の学習様式のモデルを提唱し，行動実験および脳画像解析によって実証したものであり，脳機能の解明において非常に重要な知見を示したものである．したがって，博士（理学）の学位に値するものと認められる．