平成十六年度
電気通信大学 大学院 修士論文

強化学習型タスクにおける
人間の行動決定に関する研究

電気通信大学 大学院
情報システム学研究科
情報ネットワーク学専攻

０３５１０２２　　佐々木 隆宏

指導教員
阪口 豊
出澤 正徳
小宮山 摂

提出年月日
平成十七年三月十日

目次

第 1 章

序論

1.1 研究の背景および目的

私達人間は，日々の生活を送っていく上でさまざまな行動決定 (意思決定) を行なっている．それらは，単純な問題に対するものであったり，第三者を伴うものであったり，または不確実性のもとであったりと，あらゆる問題や状況に対して行なわれている．それでは，人間は何に基づいてそれらの行動決定を行なっているのであろうか．この問いに対しては，条件や状況を限定した枠組みの中で，ゲーム理論や認知科学，心理学，さらには行動経済学などの分野において，多くの研究者がその回答を模索してきた．

一方，自律エージェントやロボットなどの制御問題において，最適な行動決定を学習するための枠組みとして研究されてきた強化学習 (reinforcement learning) のメカニズムが，近年，人間をはじめとする動物の脳内に存在することが明らかになってきている [1]．強化学習とは，試行錯誤を通して適切な行動戦略を獲得するタイプの機械学習の枠組みである．強化学習における学習者であるエージェントは，報酬と呼ばれる情報を手がかりに，適切な行動決定ルールを自律的に学習する．

ここで，より良い行動決定ルールを学習するために，エージェントは大きく分けて，探索 (exploration) と搾取 (exploitation) の 2 種類の行動パタンを持つ．強化学習のアルゴリズムを計算機で実行する場合，これらの行動確率を決定するパラメタを，設計者が試行錯誤または経験則によって決定しているのが現状である．一方，人間はこういった行動決定を，時間や計算のコストをかけずに逐次行なっているように思われる．では，人間を強化学習のエージェントであると見立てて，強化学習で扱われるような問題を課したときに，人間はどのような行動決定を行なうのであろうか．

本研究では強化学習型のタスクとして，最適な行動決定を，状態遷移の試行錯誤によって学習するようなタスクを課す学習実験を実施し，被験者がどのような行動決定を行なっているのかを，強化学習の枠組みから明らかにすることを目的とする．

第 2 章

強化学習

この章では，強化学習の基礎理論と代表的な強化学習アルゴリズムについて，また強化学習研究の応用分野である，脳の強化学習仮説に関する研究について解説する．

2.1 強化学習とは

強化学習 (reinforcement learning) とは，環境との試行錯誤による相互作用を通して適切な行動戦略を獲得するタイプの機械学習である [2]．近年，自律エージェントや自律ロボットに関する研究が活発に行なわれている．環境との相互作用を通して自律的に学習を行なう強化学習はそれらの学習制御アルゴリズムとして注目されている．

図 2.1 に強化学習のモデルを示す．強化学習では，学習する主体であるエージェント (agent) は環境 (environment) の状態を観測し，それに応じて行動する．エージェントの行動によって環境の状態は変化し，エージェントは環境から報酬を受け取る．これらのやり取りを繰り返すことによって，最終的にエージェントは最も大きい報酬を得られる行動を選択するようになる．

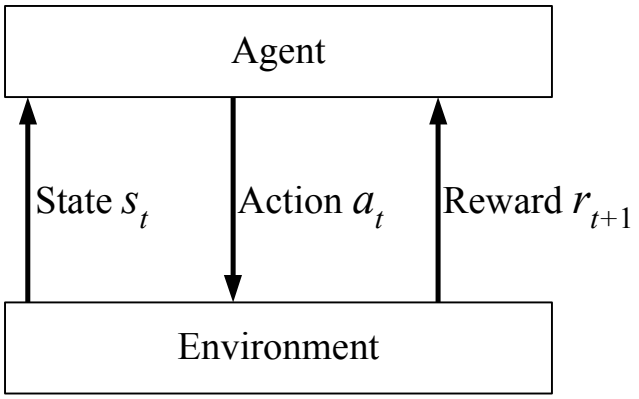


図 2.1　強化学習のモデル

2.2 強化学習の枠組み

本節では，エージェントと環境との相互作用を定義する．またエージェントの目標が割引収益の最大化であること，環境がマルコフ決定過程でモデル化できることを述べる．

2.2.1 エージェントと環境との相互作用

エージェントは環境に対して，状態の観測，行動，報酬の獲得という一連の相互作用を行なう．エージェントの目的は，これらの相互作用を通して，最終的に得られる報酬の総和を最大化する行動戦略を見つけることである．エージェントと環境との相互作用は以下のように定義できる．

1. エージェントは時刻 t において環境の状態観測 $s_t \in \mathcal{S}$(状態の有限集合) に基づいて意思決定を行ない，行動 $a_t \in \mathcal{A}$(行動の有限集合) を出力する
2. エージェントの行動により，環境は s_{t+1} へ遷移する
3. エージェントは行動の結果として環境から報酬 $r_{t+1} \in \mathfrak{R}$ を受け取る

2.2.2 強化学習の構成要素

強化学習では，以下の 3 つの要素を用いて学習が行なわれる．

- 方策 (policy)
  エージェントは観測した状態に基づき行動を選択する．この選択基準を方策という．方策は，観測した状態 $\mathcal{S}$ からその状態で取るべき行動 $\mathcal{A}$ への写像 π:$\mathcal{S} \to \mathcal{A}$ と定義される．一般的に方策は確率的であり，状態 s で行動 a を取る方策を $\pi(s, a)$ と表す．
- 報酬関数 (reward function)
  エージェントは行動した結果として環境から報酬を得る．報酬関数は環境の状態 $\mathcal{S}$ から報酬への写像 R:$\mathcal{S} \to \mathfrak{R}$ と定義される．
- 価値関数 (value function)
  価値関数は状態または行動の価値を定義する[*1]．価値とは，エージェントの現状態から将来にわたって得られる報酬の期待値である．価値関数は状態 $\mathcal{S}$ から価値への写像 V:$\mathcal{S} \to \mathfrak{R}$ と定義される．

報酬が即時的な状態の良さを表すのに対して，価値は将来にわたった長期的な状態の良さを表す．強化学習では，エージェントは最も高い価値を持つ状態へ遷移する方策を探す．

[*1] それぞれ状態価値関数，行動価値関数と呼ばれる．

2.2.3 エージェントの目的

エージェントの目的は，最終的に環境から受け取る累積的な報酬和の最大化である．
この累積報酬和は，以下の式で与えられる．

$$R_t = r_{t+1} + \gamma r_{t+2} + \gamma^2 r_{t+3} + \cdots = \sum_{k=0}^{\infty} \gamma^k r_{t+k+1} \tag{2.1}$$

ここで，γ は割引率 (discount rate)($0 \leq \gamma \leq 1$) と呼ばれるパラメタで，遠い将来の報酬ほど割り引いて評価するための値である．$\gamma = 0$ ならば，エージェントは即時報酬 r_{t+1} のみを最大化するようになり，γ の値が高いほど，エージェントは長期的に将来の報酬を考慮するようになる．

2.2.4 環境の定義

強化学習の多くの研究では，エージェントが相互作用する環境がマルコフ決定過程 (Markov Decision Process:MDP) であることを仮定している[*2]．MDP では，エージェントが状態 s_t において行動 a_t を取り，状態 s_{t+1} に遷移する確率が s_t と a_t だけによって決まる．

そのときのエージェントの状態遷移確率は確率 Pr を用いて以下のように定義される．

$$\mathcal{P}^a_{ss'} = Pr\{s_{t+1} = s' | s_t = s, a_t = a\} \tag{2.2}$$

また，エージェントが環境から受け取る期待報酬は期待値 E を用いて以下のように定義される．

$$\mathcal{R}^a_{ss'} = E\{r_{t+1} | s_t = s, a_t = a, s_{t+1} = s'\} \tag{2.3}$$

[*2] 本研究で行なった実験タスクについても，環境が MDP であることを前提として設計を行なった．

2.3 価値に基づく学習

本節では，強化学習アルゴリズムの中心である，価値の評価と最適な方策の導出方法について解説する．

2.3.1 価値関数

先に述べたように，エージェントの目的は将来的な累積報酬を最大化することであり，そのための指標となるのが価値関数である．累積報酬はエージェントがどのような行動を取るか，つまり方策 π に依存する．そのため，価値関数は特定の方策 π に関して定義される．方策 π のもとでの状態 s の価値は以下のように定義される．

状態価値関数 (state-value function)

$$V^{\pi}(s) = E_{\pi}\{R_t|s_t = s\} = E_{\pi}\{\sum_{k=0}^{\infty}\gamma^k r_{t+k+1}|s_t = s\} \tag{2.4}$$

同様に，方策 π のもとでの状態 s において行動 a を取ることの価値は以下のように定義される．

行動価値関数 (action-value function)

$$Q^{\pi}(s,a) = E_{\pi}\{R_t|s_t = s, a_t = a\} = E_{\pi}\{\sum_{k=0}^{\infty}\gamma^k r_{t+k+1}|s_t = s, a_t = a\} \tag{2.5}$$

2.3.2 価値関数の再帰性

価値関数は，特定の再帰的関係を満たす性質を持っている．

状態価値関数において，ある状態の価値と後続状態群の価値との再帰性は以下の Bellman 方程式で表される．

$$V^{\pi}(s) = \sum_{a}\pi(s,a)\sum_{s'}\mathcal{P}^{a}_{ss'}[\mathcal{R}^{a}_{ss'} + \gamma V^{\pi}(s')] \tag{2.6}$$

同様に，ある状態における行動の価値と後続状態群における行動の価値との再帰性は以下の Bellman 方程式で表される．

$$Q^{\pi}(s,a) = \sum_{s'}\mathcal{P}^{a}_{ss'}[\mathcal{R}^{a}_{ss'} + \gamma\sum_{a'}\pi(s',a')\mathcal{Q}^{\pi}(s',a')] \tag{2.7}$$

2.3.3 最適方策と最適価値関数

すべての状態 s に対して，$V^{\pi}(s) \geq V^{\pi'}(s)$ であるとき，π は π' より良い方策と定義される．他のすべての方策より良いか同じである方策は少なくとも一つ存在し，これを最適方策 π^* と呼ぶ．

最適方策は，すべての $s \in \mathcal{S}$ に対して，最適状態価値関数と呼ばれる以下の式を満たす．

$$V^*(s) = \max_{\pi} V^{\pi}(s) \tag{2.8}$$

同様に，すべての $s \in \mathcal{S}$ と $a \in \mathcal{A}(s)$ に対して，最適行動価値関数と呼ばれる以下の式を満たす．

$$Q^*(s,a) = \max_{\pi} Q^{\pi}(s,a) \tag{2.9}$$

ここで，V^* は Bellman 方程式を満たす必要があるが，この Bellman 方程式は最適方策 π^* のみに従うため，任意の方策 π に依存しない特別な形で表すことができる．これを Bellman 最適方程式と呼び，以下の式で表される．

$$V^*(s) = \max_{a} \sum_{s'} \mathcal{P}^a_{ss'}[\mathcal{R}^a_{ss'} + \gamma V^*(s')] \tag{2.10}$$

Q^* についても，同様に以下の式で表される．

$$Q^*(s,a) = \sum_{s'} \mathcal{P}^a_{ss'}[\mathcal{R}^a_{ss'} + \gamma \max_{a'} Q^*(s',a')] \tag{2.11}$$

Q^* が求まれば，最適方策 π^* は容易に求めることができる．すなわち，

$$\pi^*(s) = \arg\max_{a} Q^*(s,a) \tag{2.12}$$

以上より，Bellman 最適方程式を解くことによってエージェントの取るべき最良の行動戦略である π^* を求めることができる．しかしこの方法では，環境の状態数を $|\mathcal{S}|$ とすると $|\mathcal{S}|$ 次連立方程式を解かねばならない．$|\mathcal{S}|$ が大きい問題に対応するために，強化学習にはさまざまなアルゴリズムが提案されている．

2.4 強化学習アルゴリズム

本節では，強化学習アルゴリズムのうち代表的なものを解説する．

2.4.1 動的計画法

動的計画法 (dynamic programming) とは，小規模な問題の解を記憶しておき，より大規模な問題をそれらの解から得るという手法である．強化学習に動的計画法が適用できるのは，環境に関する完全なモデルが MDP として与えられている場合，つまり，状態遷移確率と報酬分布に関する完全な知識が前もって与えられている場合のみである．

2.4.2 TD 学習

TD 学習 (temporal difference learning) は，強化学習の中心となる代表的なアルゴリズムである．TD 学習は動的計画法とは異なり，環境の完全なモデルを必要とせず，相互作用によって得られる経験から価値関数を求めることによって学習が進行する．

最も単純な TD 学習は TD(0) と呼ばれ，時刻 t における環境との 1 回の相互作用から得られる経験を用いて，現在の状態 s_t の価値 $V(s_t)$ を以下の式によって更新する．

$$V(s_t) \leftarrow V(s_t) + \alpha\delta_t \tag{2.13}$$

$$\delta_t = r_{t+1} + \gamma V(s_{t+1}) - V(s_t) \tag{2.14}$$

ここで，α は学習率と呼ばれるパラメタで，$0 \leq \alpha \leq 1$ である．また，δ_t は，TD 誤差 (TD error) と呼ばれる．

2.4.3 Sarsa

Sarsa は，TD 学習の一種のアルゴリズムである．学習には行動価値関数を用い，状態 s_t における行動 a_t の価値 $Q(s_t, a_t)$ を，次の時刻 $t+1$ で実際に選択する行動を用いて，以下の式によって更新する．

$$Q(s_t, a_t) \leftarrow Q(s_t, a_t) + \alpha[r_{t+1} + \gamma Q(s_{t+1}, a_{t+1}) - Q(s_t, a_t)] \tag{2.15}$$

2.4.4 Actor-Critic 法

Actor-Critic 法も，TD 学習の一種のアルゴリズムである．Actor-Critic 法では，エージェント内部に価値評価部分 (critic) と行動選択部分 (actor) が独立して存在する．critic

は状態価値関数を評価し，その出力 (TD 誤差) に基づいて actor は方策を学習する．

以下は，actor の行動選択が Boltzmann 分布による softmax 手法であるときの学習則である．

$$\pi_t(s,a) = \frac{\exp(p(s,a) * \beta)}{\sum_{a'} \exp(p(s,a') * \beta)} \tag{2.16}$$

ここで，β は逆温度パラメタと呼ばれ，探索 (exploration) と搾取 (exploitation) 行動のトレードオフを決定する．また，$p(s_t, a_t)$ は，時刻 t で actor が変更可能な方策パラメタの値を表し，例えば，次式のように s_t における a_t の選択傾向の強さを更新する．

$$p(s_t, a_t) \leftarrow p(s_t, a_t) + \alpha_2 \delta_t (1 - \pi_t(s_t, a_t)) \tag{2.17}$$

ここで，α_2 は，actor の学習率であり，critic が状態価値関数を更新する際に用いられる学習率 α と区別される．

2.4.5 Q 学習

Q 学習 (Q-learning) も，TD 学習の一種のアルゴリズムである．TD 学習と同じく環境の完全なモデルを必要とせず，エージェントの経験から学習が可能である．TD 学習と異なる点は，状態の価値ではなく行動の価値の更新を行なう点である．

最も単純な Q 学習は 1 ステップ Q 学習と呼ばれ，時刻 t における環境との 1 回の相互作用から得られる経験を用いて，現在の状態 s_t における行動 a_t の価値 $Q(s_t, a_t)$ を以下の式によって更新する．

$$Q(s_t, a_t) \leftarrow Q(s_t, a_t) + \alpha[r_{t+1} + \gamma \max_a Q(s_{t+1}, a_{t+1}) - Q(s_t, a_t)] \tag{2.18}$$

Q 学習では，状態 s_{t+1} において，最大の行動価値 (Q 値) を持つ行動 a を選択している点が Sarsa とは異なる．Q 学習の方策として以下のものがよく用いられる．

- ϵ-greedy
  確率 ϵ でランダムな行動を，それ以外は最大の Q 値を持つ行動 $\arg\max_a Q(s,a)$ を選択する．
- Boltzmann rule
  状態 s において行動 a を以下の確率で選択する．

$$Pr(a|s) = \frac{\exp(Q(s,a) * \beta)}{\sum_{a'} \exp(Q(s,a') * \beta)} \tag{2.19}$$

2.5　強化学習と脳科学

近年，強化学習の計算理論が人間をはじめとした動物の脳内に存在することを示唆する研究が行なわれている．そもそも，より多くの報酬を獲得するための行動を自律的，かつ探索的に学習するという強化学習エージェントの適応的な振る舞いは，動物の行動学習の最も基本的な側面であるといえる．本節では，脳の強化学習仮説について解説する．

2.5.1　オペラント条件づけ学習

行動心理学者である Skinner は，Skinner box(図 2.2) を用いたネズミやハトの行動実験を行ない，人間をはじめとした動物の行動を説明するための基本原理として，強化 (reinforcement) による行動原理を唱えた[3]．この原理による行動学習の過程は，オペラント条件づけ学習と呼ばれる．

オペラント条件づけ学習では，動物はまずなんらかの自発的な行動 (オペラント行動) を行なう．そしてたまたま動物がある特定の行動を取ったときに餌などの報酬[*3]が与えられると，その動物は同じ行動を繰り返し行なう傾向が高まる (行動が強化される)．

このように，オペラント条件づけ学習は，試行錯誤的な強化学習アルゴリズムに近い学習であると言える．Barto が指摘するように[4]，オペラント条件づけ学習のような”trial-and-error”学習は，最も単純な形の強化学習である．

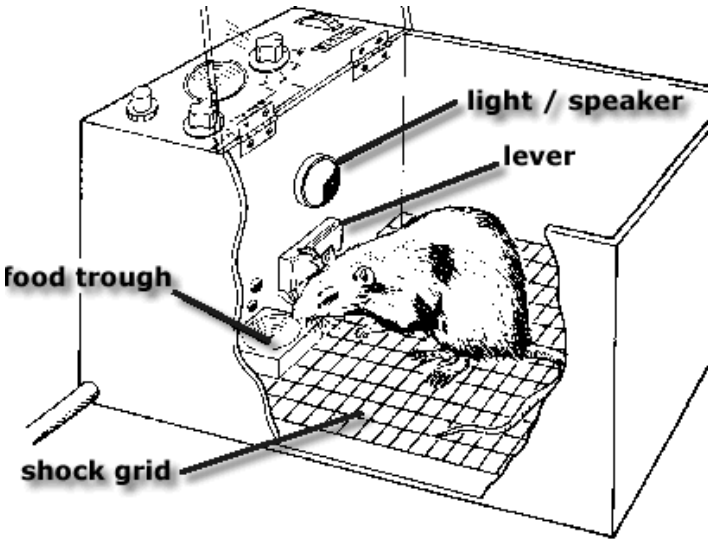


図 2.2　Skinner box．レバーを押すと餌が出てくる仕組みになっている．ネズミが偶然レバーを押すというオペラント行動によって条件づけ学習が行われる．
[出展：NIU Psychology Department website]

[*3] 心理学の分野では強化子と呼ばれる．

ただし，Skinner の実験においては，報酬である餌は，レバーを押すという単体の行動の直後に与えられる即時的なものである．これに対して，より一般的な強化学習では，入出力系列に時間的関係を持った環境を仮定している．これは，生物の場合では複数の行動列の後に，食物を獲得するなどの報酬を得るモデルである．そのため，一般的に強化学習の枠組みで捉えられているような「遅延のある報酬[*4]」とは分別して考える必要がある．

また，オペラント条件づけ学習の枠組みにおいては，報酬や罰によって動物の特定の行動が強化されるという結果についてのみ論じられており，報酬に対する内部モデルや，得られた報酬に対しての漸次的な行動選択の仕組みについては明らかにしていない．

2.5.2 脳の強化学習仮説

近年，工学的な研究の中で開発された強化学習の仕組みが，動物の脳内に存在するのではないかということを示唆する研究が行なわれている．Barto は，大脳皮質と脳幹の間に位置する大脳基底核において，報酬の期待誤差 (TD 誤差) による強化学習が行なわれているというモデルを提案した[4]．

大脳基底核は，その損傷時の症状から運動制御になんらかの形で関わることが知られていたが[5]，正常時における機能は明らかでなかった．しかし，Schultz らの研究[6]によって，大脳基底核で強く作用するドーパミンと呼ばれる神経伝達物質を放出するニューロンの活動記録から，大脳基底核が将来得られるであろう報酬予測をもとに行動を決定する上で，重要な役割を担っていることが明らかになってきた (図 2.3)．

Schultz らによると，ドーパミンニューロンは得られた報酬に反応するのではなく，予測される報酬の時間変化，すなわち強化学習で言うところの TD 誤差に反応する．設楽も，ドーパミンニューロンが報酬の期待に関与していることを示す研究を行なっている[7]．さらに，Tanaka らは，短絡的な報酬予測と長期的な報酬予測は，大脳皮質と大脳基底核を結ぶ並列回路の異なる部分で行われていることを明らかにした[8]．

また，Schultz らの研究結果を受けて，大脳基底核を Actor-Critic アーキテクチャを用いた強化学習で数理モデル化する研究が行なわれている[9, 10]．

2.5.3 学習系のメタパラメタと神経修飾物質

強化学習などの学習アルゴリズムにおいては，学習の進行を決定するパラメタが複数存在する．例えば Q 学習などでは，将来の報酬の割引率 γ，探索のランダムさを決定する逆温度 β，学習の速度係数 α などのパラメタを用いる．強化学習を最適化問題や制御問題などに適用する場合，これらのメタパラメタの設定は設計者自信が行ない，それらの最適な

[*4] 時系列上の複数回にわたる行動の結果としての報酬が，遅れて価値関数に反映されるという意味．

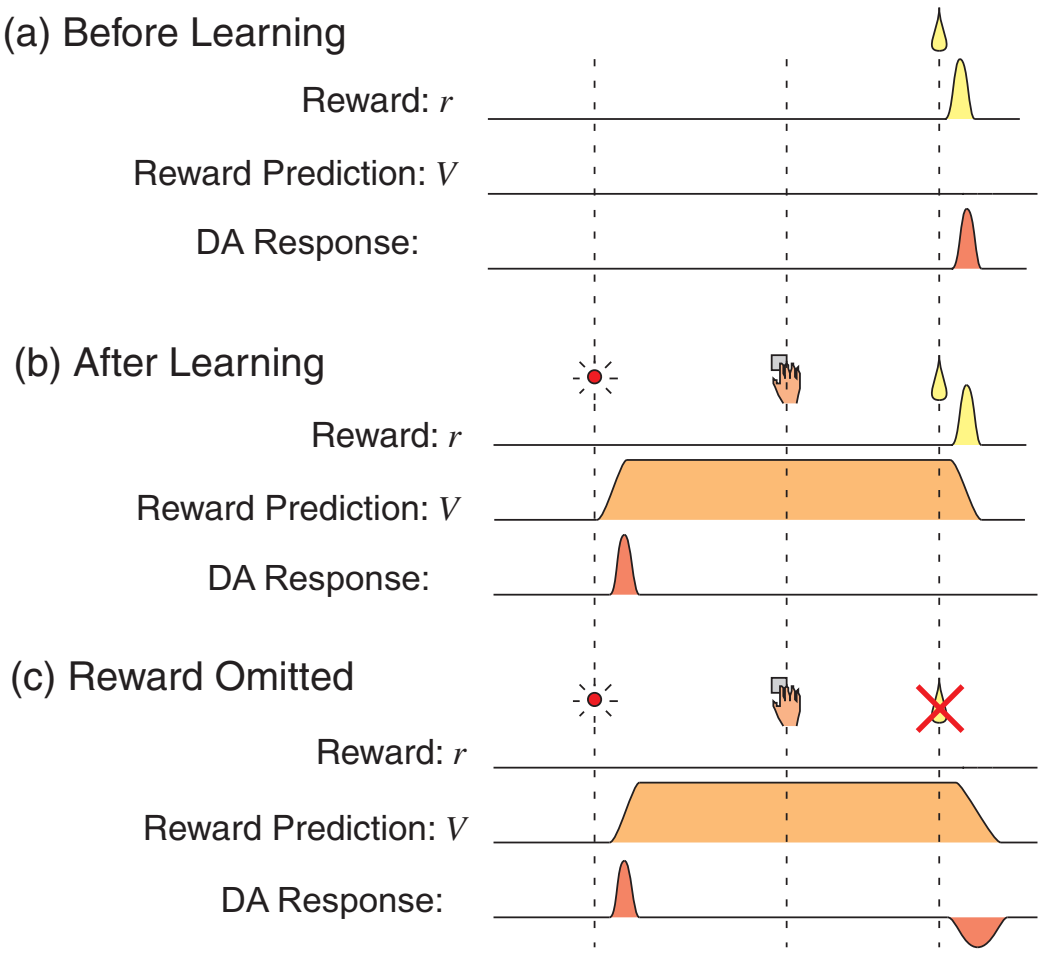


図 2.3　Schultz et al. の実験[6]．上図は，サルに，ランプがついた後，正しくレバーを押したら報酬であるジュースを与えるというタスクを課しながら，大脳基底核の一部である黒質ドーパミンニューロンの活動を記録した模式図である．ここで，横軸は時間，報酬 r と報酬予測 V の縦軸はそれらの大きさを表す．(a) 学習前，ドーパミンニューロンは突然得られた報酬そのものに対して反応する．このときの報酬予測は無し，つまり強化学習で言うところの価値は，$V = 0$ となるはずである．(b) 学習後，つまり，ランプがついたらレバーを引いて報酬が得られることがほぼ確実にできるようになると，報酬そのものに対する反応はなくなり，ランプがついたことに対する反応が大きくなる．このとき，報酬予測はランプがついた時点で立ち上がり，報酬が得られたあとにゼロとなるはずである．(c) 学習後に報酬がカットされた場合，すなわち，予測していた報酬が与えられなかった場合，ドーパミンニューロンの活動は抑制される．これらのドーパミンニューロンの活動は，報酬の予測誤差，つまり，強化学習の TD 誤差 δ_t と一致している．この実験結果は，大脳基底核は報酬予測機能を司っており，そのメカニズムが強化学習と同様であるかもしれないということを示唆している．[出展：[1]]

設定は，学習する問題の内容やその規模に依存するため，設計者が試行錯誤して決定しているのが現状である．

一方，人間をはじめとした動物は，未知の環境のもとでもさまざまな行動を自律的に学習することができる．もし脳が強化学習を行なっているとすれば，これらのメタパラメタと同様の機構が存在することが考えられるが，それらは外部の誰かにチューニングしてもらっているわけではない．このことは，脳に学習系のメタパラメタを適応的に調節する機構が存在することを示唆している．

銅谷は，自らの数理モデルと生理学的知見をもとに，メタパラメタと神経修飾物質[*5]を対応付けて考えた仮説を提案した[11]．銅谷はこの研究において，情動系への計算論的アプローチとして，強化学習におけるメタパラメタと神経修飾物質を以下のように対応付け，説明している．

- TD 誤差 δ とドーパミン
  Schultz らの実験結果による．
- 割引率 γ とセロトニン
  エージェントは，割引率が小さいほど目先の報酬のみに従った行動選択を行なう．一方，セロトニンの低下は，鬱病のような，目先の困難などにとらわれて長期的展望ができないという症状を引き起こす．
- 学習率 α とアセチルコリン
  学習率が小さいと学習の進行は遅くなり，大き過ぎると収束が不安定になる．一方，アセチルコリンは，記憶を司る海馬のシナプス可塑性の効率に影響を与えることから，何を記憶し，何を記憶しないかという学習の重み付けをしているのではないかと示唆される．
- 逆温度 β とノルアドレナリン
  エージェントは，逆温度が低いほど探索行動を，高いほど搾取行動を重視する．一方，ノルアドレナリンは，高い覚醒状態などの緊急時に多く分泌されることから，緊急時ほどリスクの低い行動決定をするためではないかと示唆される．

以上の対応付けは，仮説の段階ではあるが，脳の強化学習仮説を明らかにしていく上で注目すべきポイントの一つとして考えることができる．

[*5] ニューロンから放出される神経伝達物質のうち，脳全体に拡散的に投射され，持続的な効果を持つものの総称．

第 3 章

強化学習型タスク

本章では，本研究で行なった強化学習型タスクの学習実験の概要について，またその実験結果に対する本研究独自のアプローチである，強化学習の枠組みから捉えた解析手法について述べる．

3.1　タスクの設計

本研究の目的は，強化学習アルゴリズムの枠組みから人間の行動決定を明らかにすることである．そのため，本研究では，強化学習の枠組みから見て人間の行動決定を分析し，また，人間と計算機プログラムによる強化学習エージェントの比較を行なう．

以上の理由から，本研究では，以下のような条件を満たすタスクを題材に用いる．

- 最適な行動決定を状態遷移の試行錯誤によって学習するタスク
  1. 環境は MDP である
  2. タスクを最後まで達成してはじめて報酬が得られる
  3. 計算機の強化学習プログラムにも実行できる

本研究では，有限オートマトンの状態遷移を試行錯誤によって学習するタスクを用いた．本タスクは，上記の定義を満たしており，なおかつ以下の設計方針に従って設計されるものである．

- 被験者を強化学習エージェントであると見立てて学習を行なえるもの
  本研究では，被験者を強化学習型タスクの環境に配し，強化学習エージェントであると見立ててその行動決定を明らかにするという手法を用いる．
- 強化学習のアルゴリズムを用いるとうまく学習が収束するもの
  強化学習の枠組みから人間の行動決定を観察するためには，被験者を強化学習エージェントと見立てたとき，強化学習のアルゴリズムに則って行動決定を行なうことが

適切な戦略であるようなタスクが望ましい.

- 被験者の学習できるレベルのもの
  計算機プログラムにとっては問題がなくても，状態空間が広大すぎるなどの理由で，被験者の学習が実質的に不可能なタスクであると比較が行なえない.
- 学習に前もった知識が有効でないもの
  被験者は試行錯誤により状態遷移を学習するので，被験者が既に獲得している記憶や知識が学習に有効に働かないタスクである必要がある.
- 状態や行動に認知的な意味づけがされにくいもの
  被験者は強化学習エージェントとして学習に参加するので，状態やその状態における行動といった強化学習に必須の要素に，被験者が独自に認知的意味づけをしてしまうことによって，強化学習の枠組みから逸れた学習をしてしまうことを避ける.

以下に，本実験で用いた強化学習型タスクである，有限オートマトンの状態遷移を試行錯誤によって学習するタスクの概念図を示す (図 3.1)[*1].

本タスクには初期状態とゴール状態が存在し，タスクの開始時，被験者は初期状態にいる．被験者は 2 種類の行動を選択することによって状態遷移を繰り返しながら，目的であるゴール状態への遷移を目指す．この一連の区切りをエピソードと呼ぶ．被験者はこのエピソードの繰り返しによって，環境に関する知識 (どの状態でどの行動を行なうとどの状態へ遷移するのか) を学習していく.

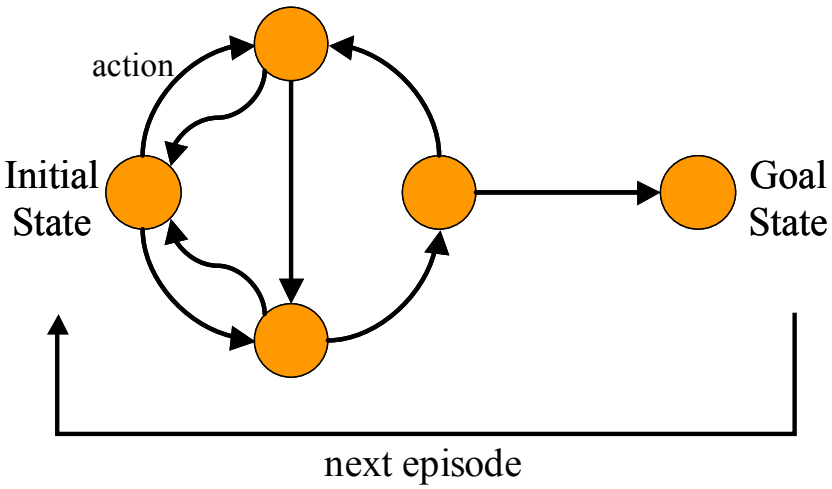


図 3.1　有限オートマトンの状態遷移の概念図 (状態数 5)

[*1] オートマトンが自分自身へ遷移しない仕組みであるのは，後述する報酬との連関による.

3.2　実験課題の概要

本節では，前節で決定された設計方針に基づいて設計されたタスクの概要を示す．

3.2.1　実験環境

前節の設計方針に基づき，強化学習型タスクとして，モニタとマウスを用いて有限オートマトンの状態遷移を試行錯誤により学習するタスクを設計し，被験者実験を行なった(図 3.2)．モニタの右側にはオートマトンの状態としてそれぞれにユニークな色[*2]が表示され，被験者の 2 種類の行動はマウスの左右のボタンのいずれかを押下することによって実現される．ボタンが押されると状態が遷移し，被験者は状態が遷移したことを色の変化として認識する．モニタの左側には，現在エピソードの何回目であるか，また，ゴール状態に遷移したときには，その旨とともにゴールまでに要したステップ数が表示される．本実験では，エピソード数を 10 と定め，被験者へも教示している．つまり，1 つのタスクについて 10 回ゴール状態へ遷移することがそのタスクの終了条件となる．

3.2.2　被験者への指示

実験開始にあたって被験者へ与えた指示を以下に示す．なお，被験者にはタスクの最短ステップ数を教示している．最短ステップ数とは，本実験タスクにおいて，初期状態からゴール状態まで遷移するのに要する最短の行動数のことである．

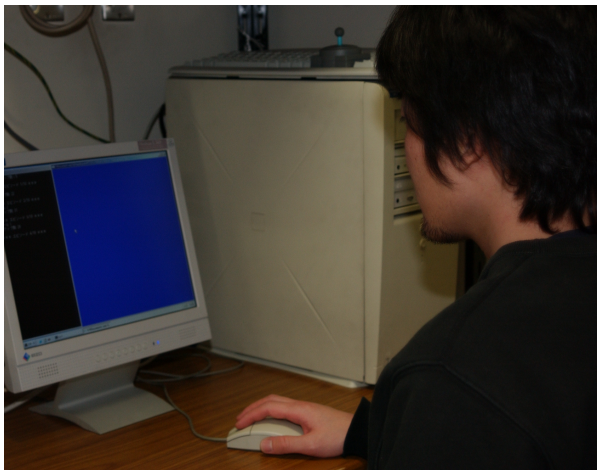


図 3.2　実験の様子

[*2] 色による状態提示としたのは，文字や記号などによる提示と比較して，状態に対する被験者の認知的意味づけを難しくさせるためである．

- なるべく少ないステップでゴールせよ
  1. ゴールまでの最短経路を発見するのが目的ではない
  2. 10 回のエピソードで要したステップの合計が最小になるような行動決定をせよ
  3. 最短ステップ数でゴールできる行動系列を発見したらそれを繰り返して実行せよ
- 行動決定に要する時間は問わない (好きなタイミングでボタンを押して構わない)

強化学習型タスクである本実験タスクにおける行動は，大きく探索 (exploration) 行動と搾取 (exploitation) 行動の 2 種類に分別することが出来る．探索行動は，文字通り試行錯誤によって環境に関する知識を獲得する行動であり，搾取行動とは，獲得した知識に基づいて行なう行動を意味する．ここで，上記 1，2 のような指示を与えたのは，被験者の探索行動をある程度制限するためである．

3.2.3　タスクのトポロジーと最短正解系列

本研究では，有限オートマトンの状態遷移を学習するタスクをトポロジー別に 3 種類用意し，それぞれに 10 人ずつの被験者を用意して実験を行なった (表 3.1)．

ここで，同一実験内の各タスクにおけるオートマトンのトポロジーについては，タスク間の比較を行なうために，すべて同一であるように設計されている．付録 A.1，A.2，A.3 に，それぞれの実験におけるタスクのトポロジーを示す．被験者は実験において，ランダムな順序で全種類のタスクすべてを実行するが，これらのタスク間では，状態を表現している色の配置[*3]と 2 種類の行動の配置が変更されて設計されている．従って，トポロジーは同一であっても，異なるタスク間において環境の知識は流用できず，被験者にとっては異なる課題であるように感じる．

また，実験 1，2，3 におけるそれぞれのタスクには，それぞれ初期状態から 5，7，9 ステップでゴールに辿り着くことができる最短の経路が 2 種類ずつ存在する[*4]．初期状態からそれらの最短経路を遷移してゴールに辿り着くまでの行動列は 2 種類存在し，これらの行動列を最短正解系列と呼ぶ．最短正解系列は，例えば L-L-R-L-L(左ボタンを 2 回，右

	最短ステップ数	状態数	タスク数
実験 1	5	11	10
実験 2	7	13	10
実験 3	9	15	8

表 3.1　トポロジーの違いによって分別された 3 種類の実験

[*3] 初期状態のみ赤色で固定されている．

[*4] 最短ステップ数は，人間の記憶のマジックナンバーが 7±2 であるとされていることより決定された．

ボタンを 1 回，左ボタンを 2 回押す行動系列) などである．

ここで，タスクが持つ 2 種類の最短正解系列は，それぞれ行動のスイッチ回数が異なるように設計されている．行動のスイッチ回数とは，行動系列の実行中に，右左のボタンの切り替えが何回存在するかというもので，上述の例においては，最初の 2 つの行動は左ボタンの行動列であり，3 つ目の行動で右ボタンへスイッチする．さらにその次の行動で左ボタンの行動列へとスイッチするので，スイッチ回数は 2 となる．例えば，実験 1 においては，それぞれのタスクは表 3.2 に示されるような最短正解系列を有している．

	最短正解系列	スイッチ回数
タスク a	L-L-L-L-R	1
	R-R-L-L-R	2
タスク b	R-R-L-L-L	1
	L-L-R-L-L	2
タスク c	L-R-R-R-R	1
	R-L-L-R-R	2
タスク d	L-R-L-R-L	4
	R-L-R-R-L	3
タスク e	L-L-L-L-L	0
	R-L-R-L-L	3
タスク f	L-R-R-R-L	2
	R-L-L-R-L	3
タスク g	R-L-L-R-R	2
	L-L-R-R-R	1
タスク h	L-R-R-L-L	2
	R-R-R-L-L	1
タスク i	L-R-L-L-R	3
	R-L-R-L-R	4
タスク j	L-R-L-R-R	3
	R-R-R-R-R	0

表 3.2　実験 1 における各タスクの最短正解系列とスイッチ回数 (最短正解系列の L はマウスの左ボタン，R は右ボタンを押す行動を示す．)

3.2.4 報酬

一般的に，報酬とは，労働の対価としての金銭や物品という意味で扱われることが多い．しかし，強化学習の枠組みにおいては，報酬は強化因子 (reinforcer) という概念で用いられる．強化学習において，報酬は対価であると同時に，未来の行動決定に影響を及ぼすものである．報酬は教師あり学習の正解とは異なるが，合目的的な行動の結果として与えられる．

本実験タスクにおいては，被験者の目的は試行錯誤を伴う行動決定によって状態遷移を繰り返し，ゴール状態に辿り着くことである．よって，ゴール状態への遷移を，被験者にとっての報酬であるとみなしている．ここで，このような報酬という要素の存在する強化学習型の系列学習タスクを設計する際に，設計者が気を付けなければならないのは，被験者にとっての報酬と設計者が意図した報酬は，必ずしも一致しない可能性があるという点である．

本実験タスクにおいても，エピソードを終えてゴールに辿り着くしたことだけが，その被験者にとっての報酬であるとは限らないかもしれない．これは，被験者がゴールに辿り着くこと以外の要素を報酬として感じている場合などである．例えば，なかなかゴールに辿り着けない時間が長い間続き，被験者がタスクの学習に面白みを感じないとき，被験者にとって，一つ一つの行動に伴う状態遷移は負の報酬であると言えるかもしれない[*5]．

また，ゴールに辿り着くことが報酬であったとしても，その大きさや，なぜ被験者がそれを報酬と捉えているかの理由については，タスクの設計者の意図を超越したものである可能性がある．例えば，被験者が難しいと感じたタスクをゴールしたときの喜びは，おそらく簡単であると感じたタスクのものより大きいだろう．また，被験者が，それまでのエピソードでの記憶を生かし，自信を持って行動決定したときのゴールと，ランダムな行動決定をしていてたまたま辿り着いた際のゴールでは，報酬の大きさは異なるかもしれない．さらに，被験者は複数のタスクを連続で実行するので，報酬の大きさは学習タスクへの慣れや飽きといった，被験者の内面的なモチベーションにも依存するであろう．

このように，報酬とは，本来人間の内部で生成され，解釈されるものであると考える．しかし，報酬は強化学習の枠組みに必須の要素であり，後述するように，本実験タスクを強化学習の枠組みから捉えた解析においては，ゴール状態に遷移する行動に正の報酬を与えている．

[*5] 本実験タスクにおいて，オートマトンが自分自身に遷移しない設計としたのは，状態が変化しないことによる報酬バランスの崩れを防ぐためである．行動を行なっても状態が遷移しない (同じ状態に遷移する) 環境と比較して，いずれの行動によっても状態が遷移する環境は，状態遷移すること自体が被験者にとっての隠れた報酬となってしまう可能性は低くなると考えられる．

3.3　強化学習の枠組みから捉えた解析

本節では，人間の行動決定について，強化学習の枠組みから捉えるための解析手法について述べる．

3.3.1　人間を Q 学習エージェントと見なした解析手法

本研究では，人間である被験者を，強化学習のエージェントであると見立てることによって，その行動決定を強化学習アルゴリズムの観点から明らかにすることを試みる．本研究では，被験者が Q 学習のエージェントであると見立てて解析を行なった．

ここで，強化学習のアルゴリズムに Q 学習を用いたのは，Q 学習は強化学習の代表的なアルゴリズムの一つであり，また行動価値 (Q 値) を明示的に扱っているためである．そのため，本実験タスクような，ある状態における 2 種類の行動の，相対的な価値の大きさを学習していくような戦略が合理的であると考えられるような環境において，状態価値ではなく，ある状態における行動決定の指針となる Q 値を明示的に扱っていることによって，よりミクロな視点から人間の行動決定を解析できる．

3.3.2　ゴールにおける報酬

本実験タスクにおいて，エージェントはゴール状態に遷移することによって，唯一の報酬である正の値 1 を獲得する．

3.3.3　適正度の履歴を取り入れた Q(λ) 学習

多くの強化学習研究では，状態数の多い環境が扱われることが多く，また，非常に多くのエピソード数を要して学習を行なう．これに対して，本実験タスクの特徴として，状態数が少なく，またエピソード数も極めて少ない (10 回) という点が挙げられる．このような環境において，通常の 1 ステップ Q 学習[*6]を用いて，実行された行動に関する分析を逐次行なうことは困難である．なぜなら，Q 学習アルゴリズムにおいて，Q 値の更新が行なわれるのはエピソードごとであり，ゴール行動にのみ報酬が与えられる設定においては，報酬に遅延が存在する．つまり，ゴールから遠い状態における行動の Q 値が値を持つためには，エージェントは多数のエピソードを経験する必要がある．そのため，Q(0) 学習アルゴリズムを用いて Q 値による解析を行なおうとすると，エピソードの回数が少ない本実験タスクにおいては，解析できない状態または行動が多くなってしまう．

[*6] Q(0) 学習とも呼ばれる．

このような問題を回避する方法の一つとして，行動ごとに罰 (punishment) である負の報酬を与えるという方法が考えられる．しかし，3.2.4 項で述べたように，行動ごとの報酬の絶対的な大きさをタスクの設計者が決定することは難しく，本研究ではゴール行動にのみ報酬を与える方法を用いた．

そこで本研究における強化学習の枠組みから捉えた解析では，Q(0) 学習とともに，Q 学習を拡張したアルゴリズムである，Q(λ) 学習を解析に用いた．Q(λ) 学習では，適正度の履歴 (eligibility trace) と呼ばれる，エージェントの経験した状態または行動の一時的な記録を用いる．適正度の履歴は，各状態または行動が過去にどれだけ訪問，実行されたかを表していて，この値が高いほど，Q 値の更新に大きく影響を与える．Q(0) 学習においては，Q 値の更新がただ 1 つの状態 s_t についてだけであるのに対し，Q(λ) 学習では，適正度を持ったすべての状態，行動群について Q 値の更新が行なわれる．

具体的には，本研究では，Singh の置換履歴 (replacing trace) [12] を用いた Q(λ) 学習を，被験者および計算機の Q 学習エージェントの解析に用いた．ここで，適正度 e および Q 値の更新規則は，以下のようになる．

$$e_{t+1}(s) = \begin{cases} \gamma\lambda e_t(s) & s \neq s_t \\ 1 & s = s_t \end{cases} \tag{3.1}$$

$$\delta_t = r_{t+1} + \gamma \max_{a'} Q(s_{t+1}, a') - Q(s_t, a_t) \tag{3.2}$$

$$Q(s, a) \leftarrow Q(s, a) + \alpha\delta_t e_t(s, a) \tag{3.3}$$

ここで，λ は適正度の履歴を減衰させるパラメタで，$0 \leq \lambda \leq 1$ である．

3.3.4 Q 値の差分による合理性の判定

本研究の強化学習の枠組みから捉えた解析におけるアプローチとして，ある状態において行なった行動の Q 値から，もう一方の行動の Q 値を差し引いた，Q 値の差分 (図 3.3) と呼ぶ値を用いたものが挙げられる．本実験タスクにおいて，ある状態における行動の種類は 2 種類であり，それぞれに Q 値を持つ．ここで，Q 値とはもともと，その行動の価値を表すものである．よって，2 種類の行動のうち，決定した行動がどれくらい合理的であったかという指標に，その行動の Q 値の差分を用いて解析を行なった．これにより，行なった行動の合理性の大きさを定量的に表すことができる．

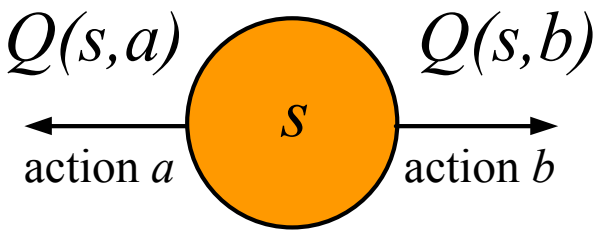


図 3.3　Q 値の差分：状態 s における行動 a の Q 値の差分を，$Q(s,a) - Q(s,b)$ とする．

3.3.5　メタパラメタの設定

本研究における強化学習の枠組みから捉えた解析において，Q(λ) 学習のメタパラメタは一律に以下の数値を用いた．

- 学習率 $\alpha = 0.1$
- 割引率 $\gamma = 0.9$
- eligibility trace の減衰率 $\lambda = 0.9$

第 4 章

解析結果とその考察

本章では，被験者へ実施した強化学習型タスクの学習実験による結果と，その考察について示す．本研究では，さまざまな視点から実験タスクにおける人間の行動決定について解析を行なった．これらについて，解析の立場により，3 つの節に分別して示す．

4.1　学習結果とその全体的な分析

本節では，学習実験による結果を，学習曲線やタスクの遂行に要したステップ数などのデータから分析した結果を示す．

4.1.1　学習曲線

各タスクにおける全被験者の学習の収束の様子を調査するため，実験 1 におけるタスク a〜j の全 10 種類のタスクにおける学習曲線を求めた (図 4.1〜図 4.10)．

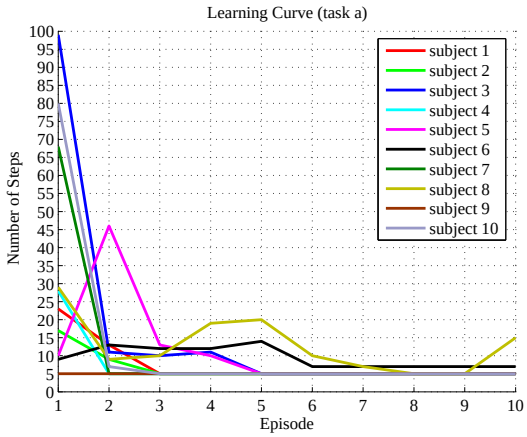


図 4.1　実験 1 タスク a の学習曲線

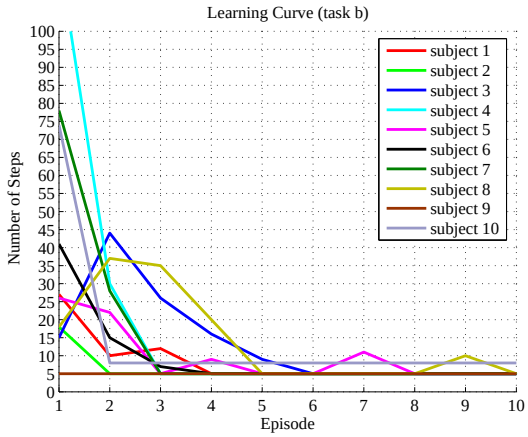


図 4.2　実験 1 タスク b の学習曲線

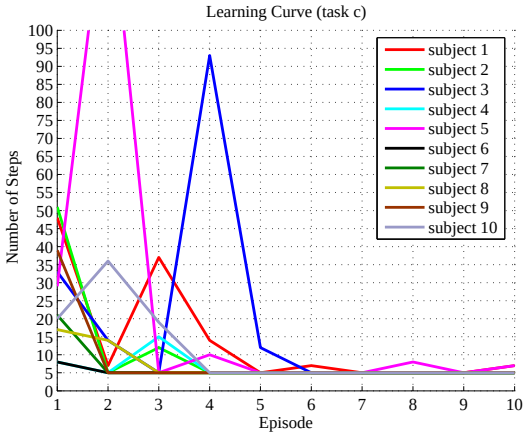


図 4.3　実験 1 タスク c の学習曲線

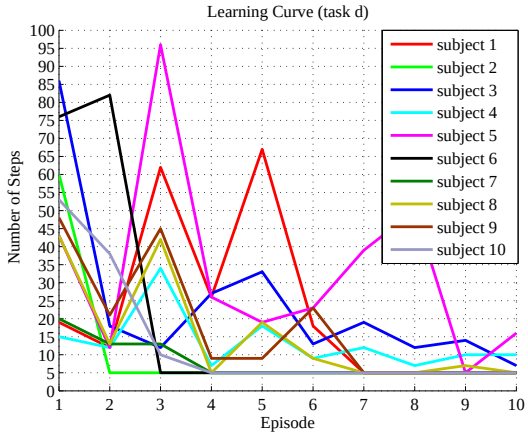


図 4.4　実験 1 タスク d の学習曲線

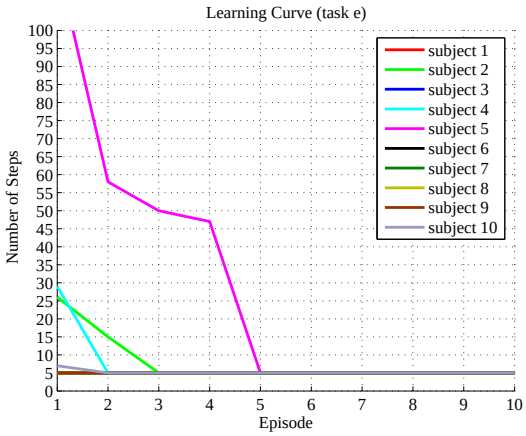


図 4.5　実験 1 タスク e の学習曲線

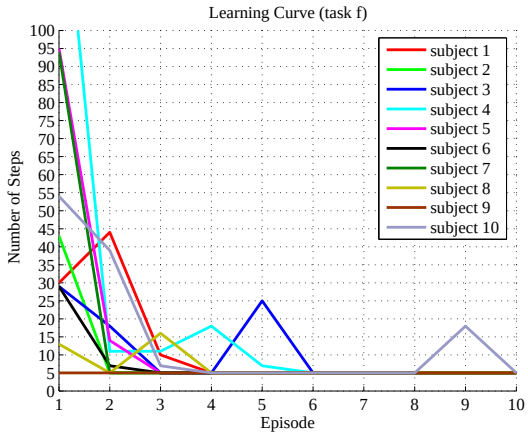


図 4.6　実験 1 タスク f の学習曲線

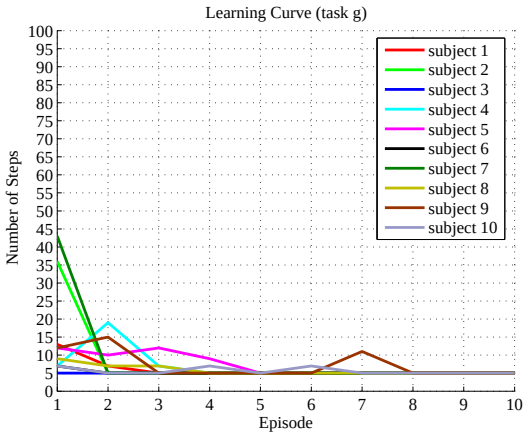


図 4.7　実験 1 タスク g の学習曲線

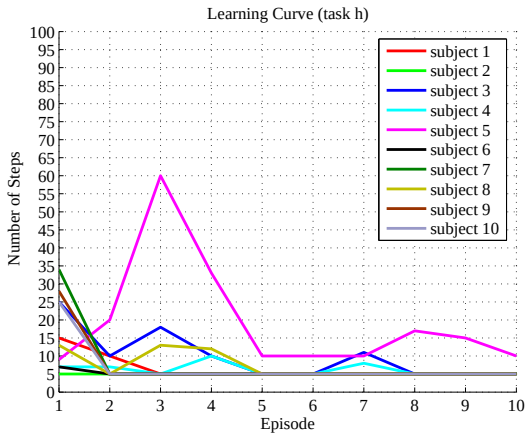


図 4.8　実験 1 タスク h の学習曲線

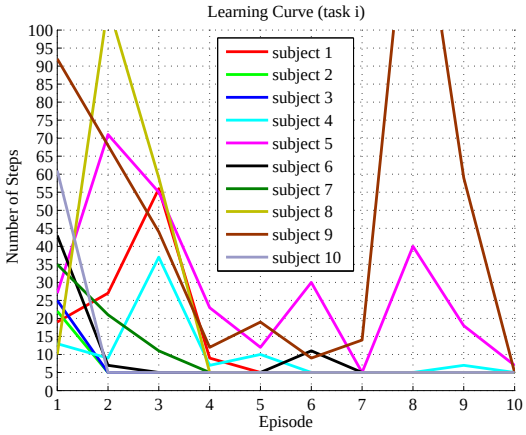


図 4.9　実験 1 タスク i の学習曲線

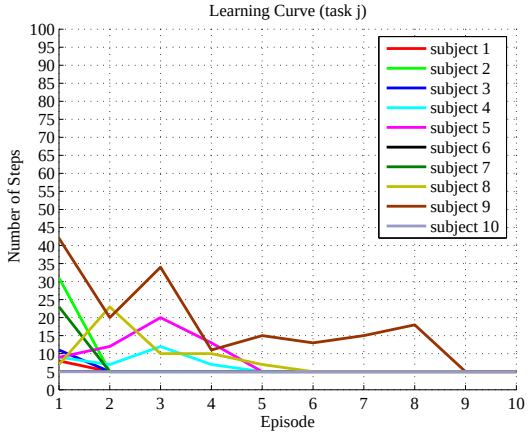


図 4.10　実験 1 タスク j の学習曲線

結果 1

以上の結果から，実験 1 におけるすべてのタスクにおいて，多くの被験者について学習の収束傾向が見られた．また，収束した学習曲線が後に発散してしまう例も見られた．

考察 1

多くの被験者について学習の収束傾向が見られたことについては，被験者がより少ないステップ数でゴールに辿り着くという目標の下で，オートマトンの状態遷移を学習し，タスクを遂行することができたということを示している．同時に，実験 1 の 10 種類のタスクが，被験者にとって学習できるレベルの問題であったことを示している．

実験 1 のタスクにおいて，初期状態からゴールまでの最短ステップ数は 5 である．そのため，初期状態から 5 ステップでゴールに辿り着くことが，これらのタスクにおいて最も学習が収束した状態であるということができる．

図 4.1 より，タスク a において，被験者 5 はエピソードの 6 回目から最終エピソードまで，初期状態から 7 ステップを要する行動を選択している．また，これらの行動はすべて同一の正解系列であった．被験者にはタスクの最短ステップ数が 5 であることを教示しているため，さらに探索行動を継続すれば最短系列を発見できる可能性があることが既知である．このような例については，この被験者は探索行動を取る選択を捨て，搾取の行動を選択し続けたと推察することができる．本実験では被験者への指示として，被験者の探索行動をある程度制限すべく，全エピソードを通してのステップ数が最短となるような行動戦略を求めている．このような目的を持った強化学習型タスクにおいて，被験者は残りのエピソード数を考慮しつつ探索-搾取の行動戦略を決定し，残りエピソード数が少なくな

るにつれて搾取行動を多く取ることが知られており[13]，以上の例における被験者の行動は，この結果を支持するものであると考えることができる．

また，図 4.9 より，タスク i において，被験者 9 は，エピソードの 7 回目までは学習が収束傾向にあるが，エピソードの 8，9 回目で要したステップ数が大きく増加している．このように，学習曲線が収束しつつある後に発散してしまっている例については，大きく 2 種類に分けてその原因を推察することができる．

- 被験者が探索行動を行なっていたと仮定した場合
  それまでのエピソードで学習しているはずの，環境に関する知識 (状態遷移の仕組み) が，より少ないステップ数でゴールするという目的において生かされていない．またはその知識の記憶が定かでなく，誤った記憶による行動，またはランダムな行動を取り，結果的により大きいステップ数を要してしまっている．
- 被験者が搾取行動を行なっていたと仮定した場合
  被験者が以前のエピソードで行動した正解系列の記憶を忘却している．

本実験においては，被験者が決定した行動について，それが被験者にとって探索行動であるか搾取行動であるかといったインタビューを，行動ごとには逐一行なっていない．また，被験者が常に自分の行動について，それが探索行動であるか搾取行動であるかのどちらに分類されるものであるかを意識して行動決定しているとは限らない．ゆえに，以上の場合分けは明確に特定できるものではなく，特定したとしても，その行動決定を行なった時点での被験者の環境に関する知識の正確さを確認することは難しい．よって，上記の例のような学習曲線に表れる学習の発散については，それぞれの原因が複合して影響していると考えることが妥当である．また，いずれの場合においても，単純なマウスの押し間違いのような，意図しない運動によるミスが原因で，学習曲線が発散している可能性も考えられる．

また，タスク間の学習収束の差を相対的に比較するため，学習曲線の傾きが全体的に落ち着いて収束していると思われるタスク a，タスク b と，収束が全体的に遅いと思われるタスク d，タスク i に注目して，それら 4 種類のタスクにおける 10 人の被験者の平均学習曲線と，標準偏差を求めた (図 4.11〜図 4.14)．

結果 2

図 4.11 と図 4.12 より，タスク a やタスク b では平均学習曲線がなだらかに推移しており，エピソードの増加に伴う被験者ごとのばらつきが小さい傾向が確認できる．これに対して，図 4.13 と図 4.14 より，タスク d やタスク i では，タスク a やタスク b と比較して，収束が遅く，被験者ごとのばらつきが大きい傾向が確認できる．

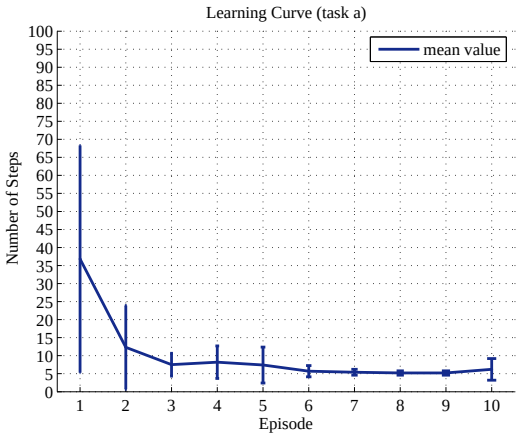


図 4.11　実験 1 タスク a の平均学習曲線

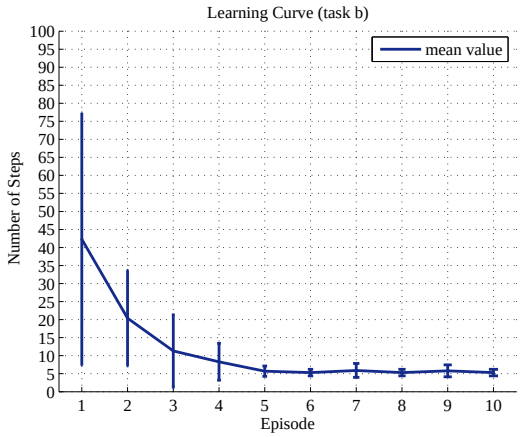


図 4.12　実験 1 タスク b の平均学習曲線

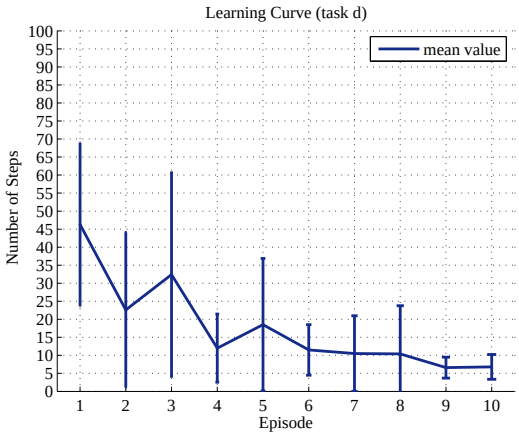


図 4.13　実験 1 タスク d の平均学習曲線

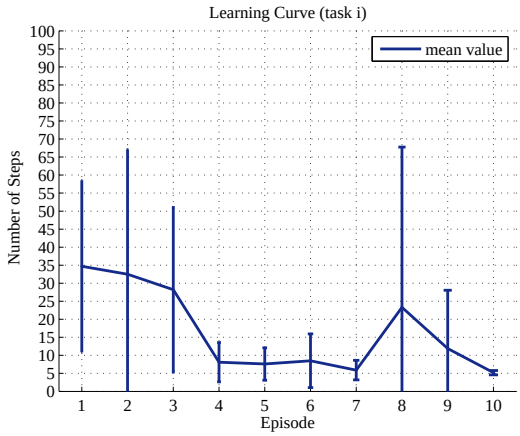


図 4.14　実験 1 タスク i の平均学習曲線

考察 2

これらの平均学習曲線に関する分析結果により，実験 1 において，同一のトポロジーであるタスク間に学習難易度の差が存在することが示された．

さらに，実験 1，2，3 の平均学習曲線を求めた (図 4.15，図 4.16，図 4.17)．

結果 3

図 4.16，図 4.17 より，実験 2 および 3 においても，多くの被験者において学習の収束傾向が確認できた．

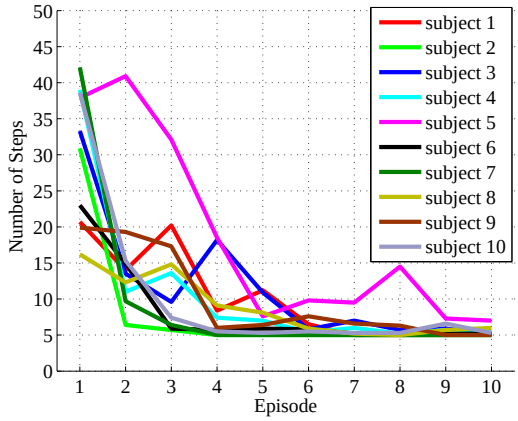


図 4.15　実験 1 の平均学習曲線

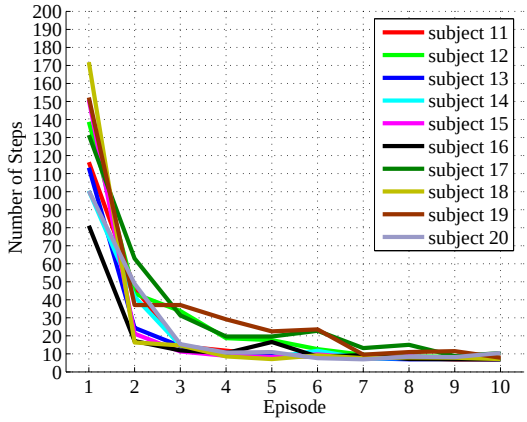


図 4.16　実験 2 の平均学習曲線

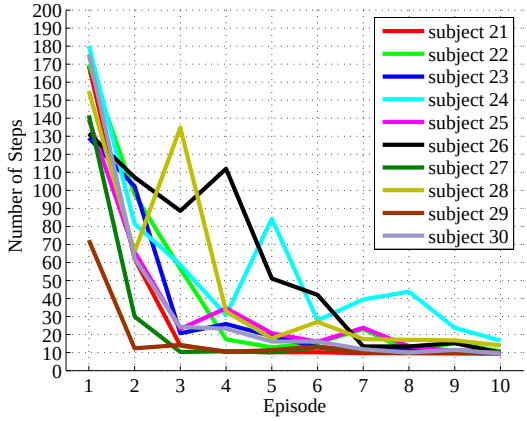


図 4.17　実験 3 の平均学習曲線

4.1.2　タスク実行に要した平均ステップ数による分析

被験者ごとのタスクの学習結果を比較するために，各タスクの遂行に要した平均ステップ数による分析を行なった．以下に，実験 1 において，各被験者が全 10 種類のタスクすべてを遂行したときに，1 つのタスクを遂行するのに要したステップ数の平均と，各タスク間におけるそれぞれのタスクの遂行に要したステップ数の標準偏差を示す (図 4.18).

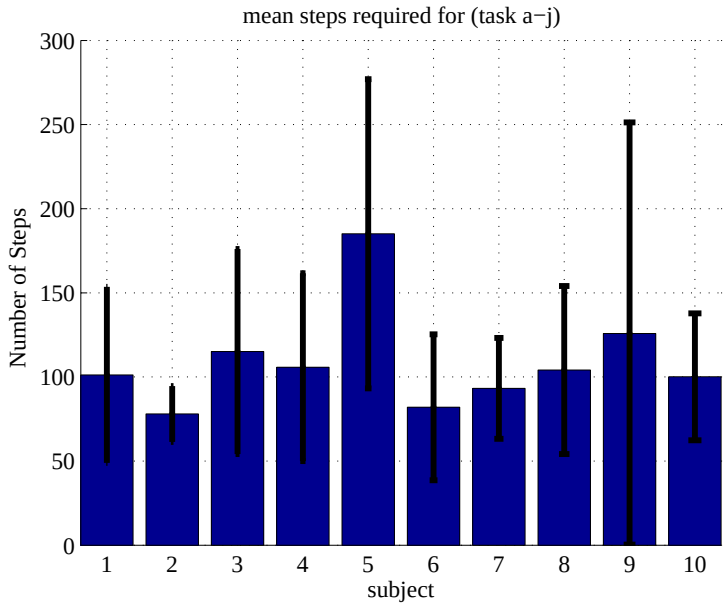


図 4.18　実験 1　被験者間の学習結果の差

結果 1

被験者間に，図 4.18 で示すような学習結果の差が見られた．最も平均ステップ数の少ない被験者 2 と最も多い被験者 5 の間には 2 倍以上のステップ数の差が見られた.

考察 1

本実験では，被験者の目的として，各タスクごとの遂行に要する総ステップ数の最小化を指示している．被験者の目的は最短の正解系列を発見することではなく，探索のために無駄な行動を多く行なうことは目的に反する．従って，上図におけるステップ数の大小が被験者間の相対的な学習結果の差，言い換えれば本実験タスクの出来の良さを端的に表したものであると考えることができる.

実験 2 および 3 についても同様に，各被験者が全種類のタスクすべてを遂行したときに，1 つのタスクを遂行するのに要したステップ数の平均と，各タスク間におけるそれぞれのタスクの遂行に要したステップ数の標準偏差を求めた (図 4.19，図 4.20).

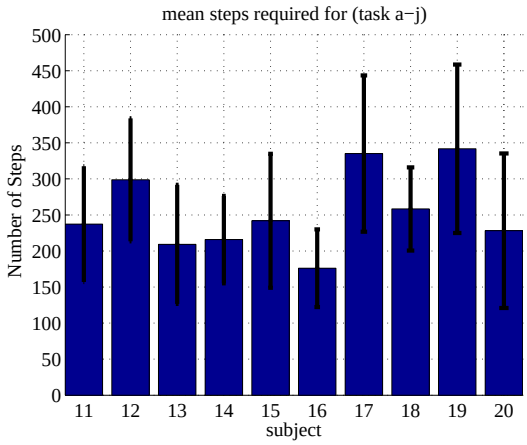


図 4.19　実験 2　被験者間の学習結果の差

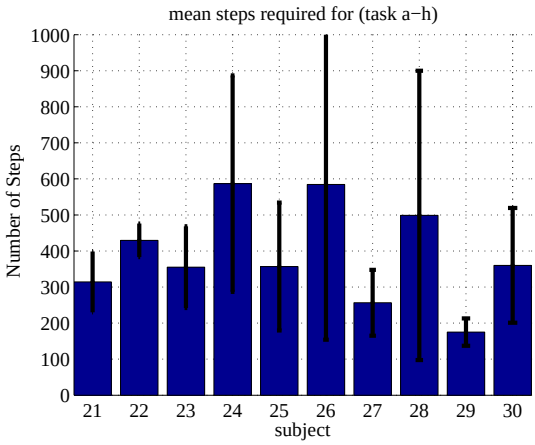


図 4.20　実験 3　被験者間の学習結果の差

結果 2

図 4.19，図 4.20 より，実験 2 および 3 においても被験者間に学習結果の差が確認された.

4.1.3　タスク間の学習難易度の分析

4.1.1 項におけるタスクごとの平均学習曲線に関する分析から，実験 1 におけるタスク間に難易度の差が存在することが示された．そこで，タスク間の学習難易度を比較するため，各タスクの遂行に要した総ステップ数を比較した．以下に，実験 1 において，全 10 人の被験者が全 10 種類のタスクを遂行するのに要したステップ数のタスクごとの平均と，被験者ごとの標準偏差を示す (図 4.21).

結果 1

図 4.21 より，タスク g はタスク遂行に要したステップ数が相対的に少なく，タスク d やタスク i はタスク遂行に要したステップ数が多いことが確認できる．最も総ステップ数の少ないタスク g と最も多いタスク d の間には 2 倍以上のステップ数の開きが見られた.

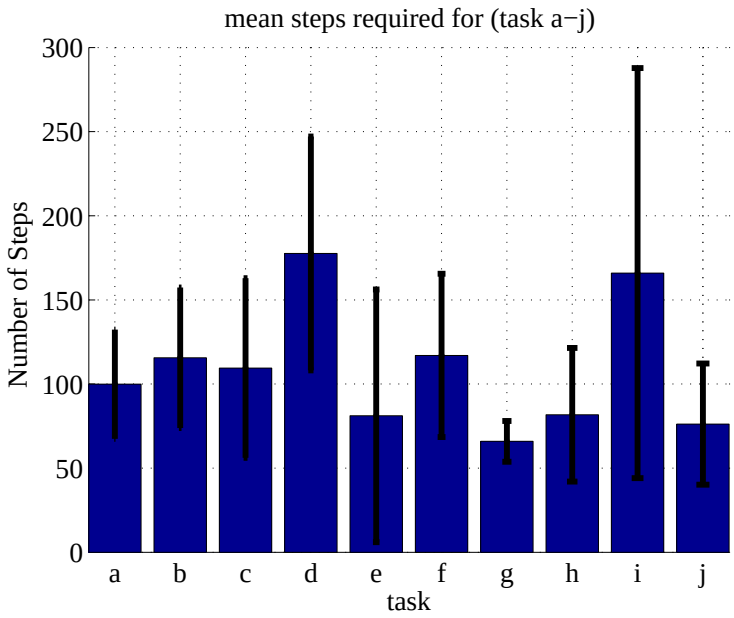


図 4.21　実験 1　タスクの学習難易度の差

考察 1

この結果より，同一トポロジーである実験 1 の 10 種類のタスク間に，相対的に学習難易度の差が存在することが示された．同時に，タスクのトポロジー以外に学習難易度を変化させる何らかの要因が存在することが明らかになった．

また，実験 2 および 3 についても同様に，それぞれのタスクの学習難易度の差を調査した (図 4.22，図 4.23)．

結果 2

図 4.22 より，実験 2 におけるタスク間には，タスク遂行に要したステップ数に大きな差が存在しないことが確認された．図 4.23 より，実験 3 において，タスク a のステップ数が相対的に多く，タスク c のステップ数が相対的に少ないことが確認された．また，タスクを実行するのに要したステップ数は，全体的に見て実験 1，2，3 の順で少ないことが確認された．

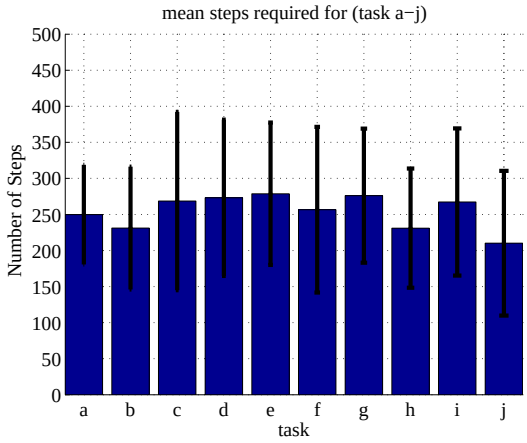


図 4.22　実験 2　タスクの学習難易度の差

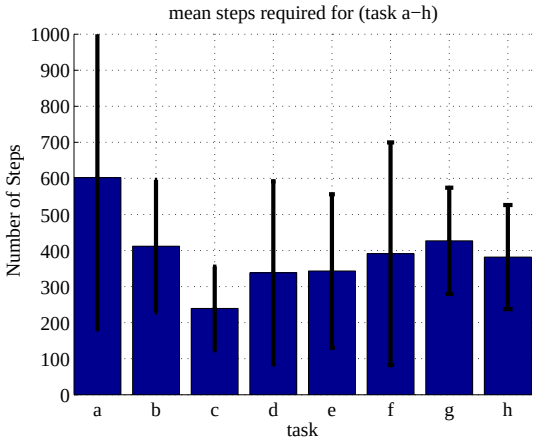


図 4.23　実験 3　タスクの学習難易度の差

考察 2

以上の結果より，実験 2 におけるタスク間には，実験 1 と比較して，大きな学習難易度の差が存在しないことが示された．実験 3 においては，タスク a の学習難易度は相対的に高く，タスク c は低いことが示された．

また，全体的なタスクの難易度としては，実験 1，2，3 の順で低いことが示された．これは，最短ステップ数が多くなるほど，またトポロジーの状態数が多くなるほど，つまりより広大な空間の探索問題になるほど，学習難易度が高くなることを裏付ける必然的な結果であると考えられる．

4.2　強化学習の枠組みから捉えた解析

本節では，強化学習の枠組みから捉えた解析として，被験者を Q 学習エージェントであると見立てたときの解析結果を示す．また，被験者をエージェントと見立てた結果と，実際の計算機プログラムの Q 学習エージェントにタスクを遂行させた結果を比較し，人間と計算機の Q 学習エージェントの行動決定の差異を探る．さらに，Q(λ) 学習と Q(0) 学習によるシミュレーション結果を比較し，どちらがより被験者の学習結果に近似しているかを分析する．

4.2.1　タスクの進行と選択した行動の Q 値の差分の遷移

本項では，被験者を Q 学習エージェントと見立ててその行動を解析した場合，一つのタスクを遂行する過程で被験者のそれぞれの行動決定がどのような Q 値の差分を持つ行動であったかをミクロに観察する．以下に，実験 1 のタスク d における被験者 1 のすべての行動をエピソードごとに分別し，それぞれの行動が持つ Q 値の差分を記録した図を示す (図 4.24).

結果

図 4.24 より，行動決定ごと (ステップごと) にその行動の Q 値の差分が増減している様子がミクロな視点から確認できる．ここで，エピソードの 1 回目において，すべての行動の Q 値の差分は 0 に等しいことが確認できる．また，エピソードの 2 回目から 6 回目までは Q 値の差分が振動しており，エピソードの 7 回目以降は安定して上昇していることが確認できる．また，エピソードの 1 回目を除くそれぞれのエピソードにおいて，一番最後の行動の Q 値の差分は正の値であり，その直前の行動の Q 値の差分と比較して大きいことが確認できる．

考察

この例において，被験者はエピソードの 7 回目からタスクの終了となる 10 回目まで，最短ステップ数である 5 ステップでゴールしている (図 4.4)．つまりこの例の場合，エピソードの 7 回目で学習は完全に収束しているということができる．

エピソードの 1 回目を見ると，すべての行動においてその Q 値の差分は 0 に等しい．エピソードの 1 回目においては，被験者は一度もゴールに到達しておらず，ゴール直前の状態において報酬が得られていない．そのため，オートマトンのすべての状態におけるす

べての行動の Q 値は 0 に等しく[*1]，結果として Q 値の差分も 0 となる．このように，エピソードの 1 回目においてすべての行動の Q 値の差分が 0 となるのは，この例に限定したことではなく，すべてのタスクに当てはまるものである．

それぞれのエピソードの最後の行動はゴールへ直接結びつく行動であり，報酬である正の値 1 を獲得できる行動である．よってその行動の Q 値の差分は正の値となる．エピソードの 2 回目以降は，ゴールに到達した際の正の報酬の獲得により，eligibility trace を持つ行動の Q 値が一斉に更新される．つまり，ゴールに辿り着くまでの，ある状態において過去に選択されたすべての行動について，その Q 値が増加する．これによって，過去に訪問歴のある状態において 2 種類のうちいずれかの行動を選択したとき，そのうちのいずれかの行動の Q 値は 0 でない値を持つ．したがって，それぞれの行動の Q 値が等しくない限り，Q 値の差分が 0 でない数値として現れてくる．エピソードの 7 回目以降は，被験者は先に述べたように最短の行動系列に対して greedy な行動決定を行なっているので，すべての行動の Q 値の差分は 0 より大きい値となる．

また，Q 値の絶対的な大きさについては，一般的にゴールからのステップ数が近い状態における行動であればあるほど大きくなる傾向がある [2]．これは，主に割引率のパラメタによる効果である．同様に，Q 値の差分は，エピソードの回数が重ねられるほど大きくなる傾向がある．これは，エピソードの増加，つまり報酬を獲得するゴール回数の増加に伴い，ゴール直前の状態における正解行動の Q 値が相対的に大きくなっていき，それに連れて伝播される Q 値も相対的に大きくなっていくためであると考えられる．

[*1] タスクの開始状態において，すべての状態におけるすべての行動の Q 値は 0 に初期化している．

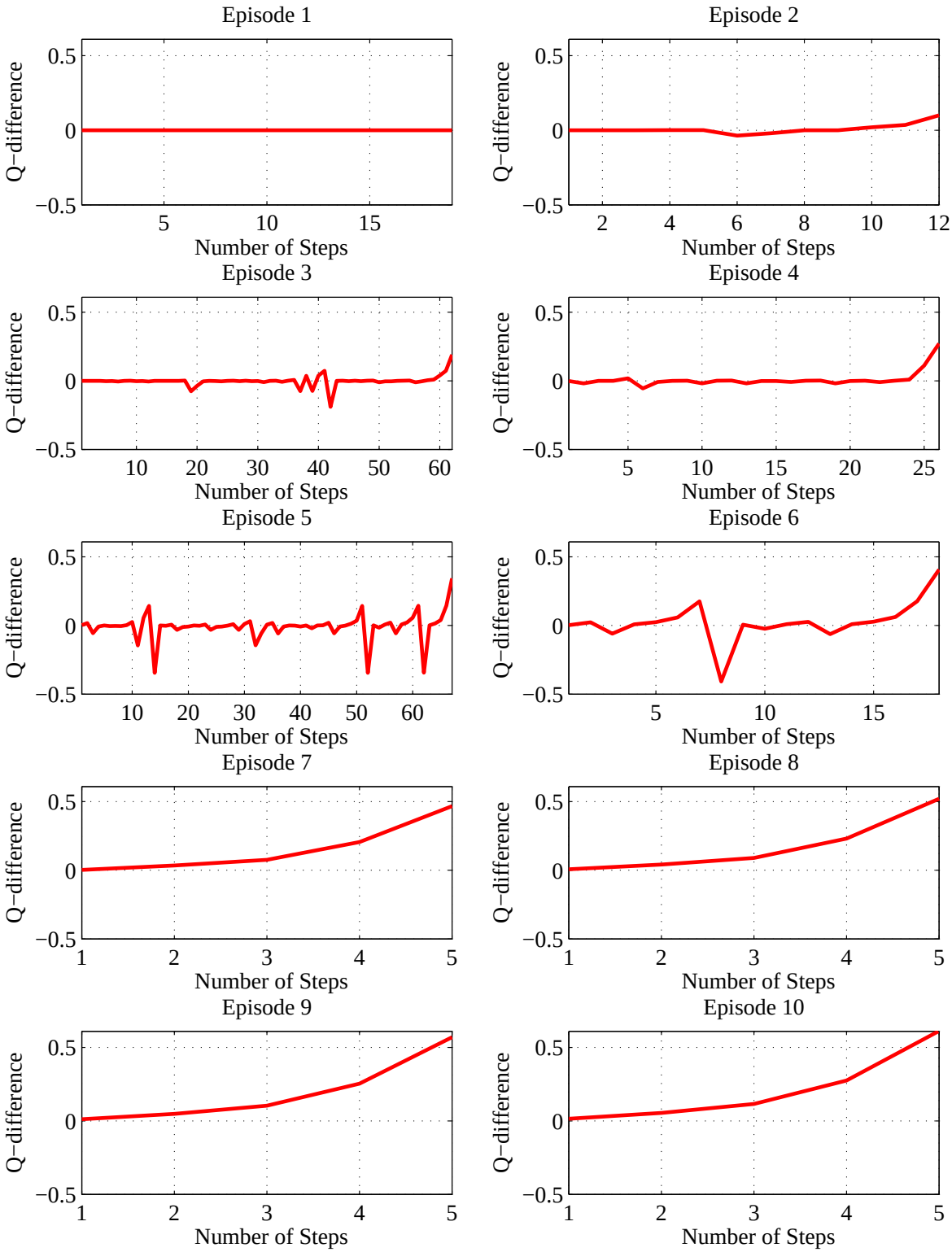


図 4.24　実験 1 タスク d における全行動とその Q 値の差分の増減 (被験者 1)

4.2.2　全タスクを通しての全行動の Q 値の差分の分布

前項では，1 タスクにおける全行動とその Q 値の差分の増減をミクロな視点で観察した．本項では，被験者の行動決定について，Q 値の差分を用いてマクロな視点から観察する．図 4.25，図 4.26 に，実験 1 の全 10 種類のタスクにおける全行動について，その Q 値の差分の大きさによって分別された行動回数の分布を示す．

結果

図 4.25 より，全員の被験者に共通する特徴として，Q 値の差分が 0 付近の行動が相対的に多いことが確認された．また，図 4.26 より，大部分の被験者について，分布の山が 0 付近を頂点として，Q 値の差分が 0 を境に大きくなるにつれてゆっくりと減少し，逆に 0 を境に小さくなるにつれて急に減少していることが確認できる．つまり，大部分の被験者について，全行動のうち Q 値の差分が 0 以上の行動が相対的に多く，特に Q 値の差分が負であり，かつその絶対値が大きい行動回数は相対的に少ないという傾向が確認された．これは，大雑把に捉えると，Q 学習の確率的行動選択基準の一つとして用いられるソフトマックス手法による振る舞いに近い．

考察

前項で述べたように，Q(λ) 学習においては，エピソード 1 回目のすべての行動，およびタスクを通して初めて訪問する状態における行動については，Q 値の差分は 0 に等しい．また，Q 値は，ゴールからのステップ数に比例して増加する割引率によって減衰するので，ゴールからの距離が相対的に遠いほどその絶対値が小さくなる傾向がある．よって，0 付近に分布している行動群は，主に以上のような状況または状態における行動であると考えられる．

また，Q 値の差分が負であり，かつその絶対値が大きい行動回数が相対的に少ない点については，学習の収束傾向と関係があると考えられる．あらゆる状態における Q 値の差分の絶対値が相対的に大きくなるエピソードの後半において，被験者の学習は収束傾向にあることが 4.1.1 項において確認されており，大きなマイナスの値を伴った Q 値の差分を持つ行動回数は，プラスのものと比較して少ないことが示された．

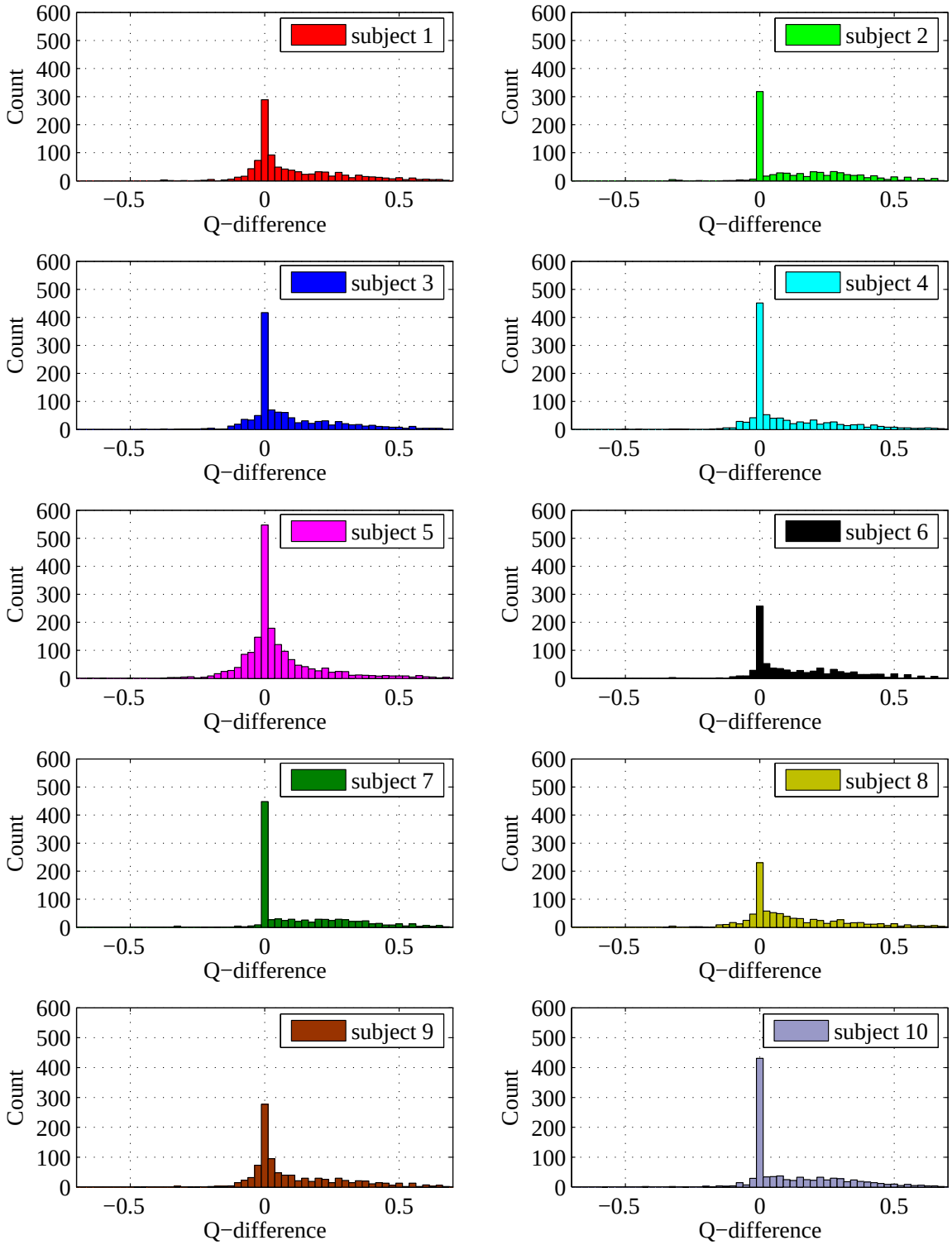


図 4.25　実験 1 の全タスクを通しての全行動の Q 値の差分

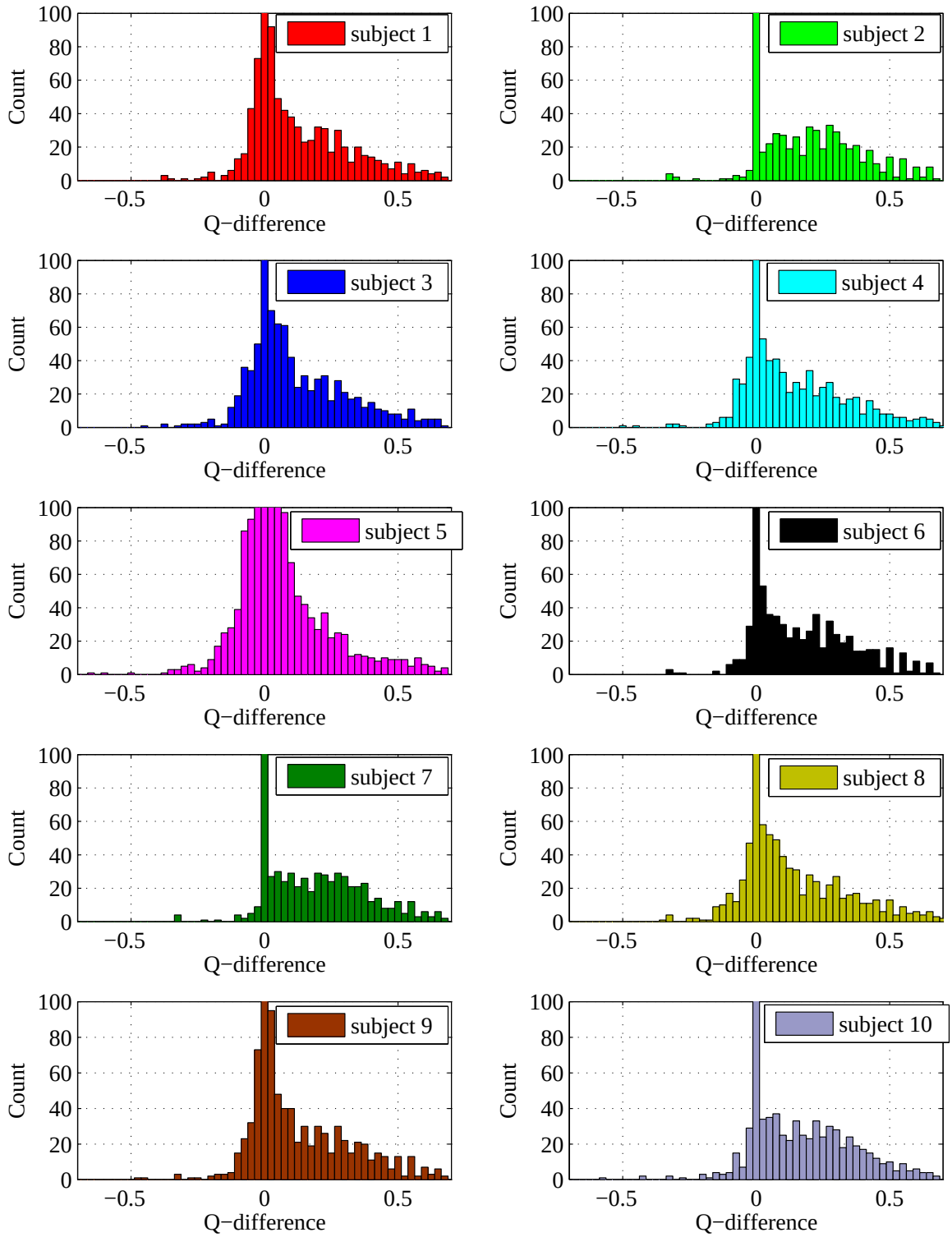


図 4.26　実験 1 の全タスクを通しての全行動の Q 値の差分 (図 4.25 の縦軸を拡大したもの)

4.2.3　全タスクを通しての全行動の Q 値の差分の内訳

本項では，引き続きマクロな視点から被験者の行動決定について探るため，被験者が実験 1 の全 10 種類のタスクにおいて行なったすべての行動を，その Q 値の差分により分類した．以下に，各被験者のすべての行動確率を，その Q 値の差分が 0 未満のもの，0 と等しいもの，0 より大きいものの 3 つのタイプに分類した図を示す (図 4.27).

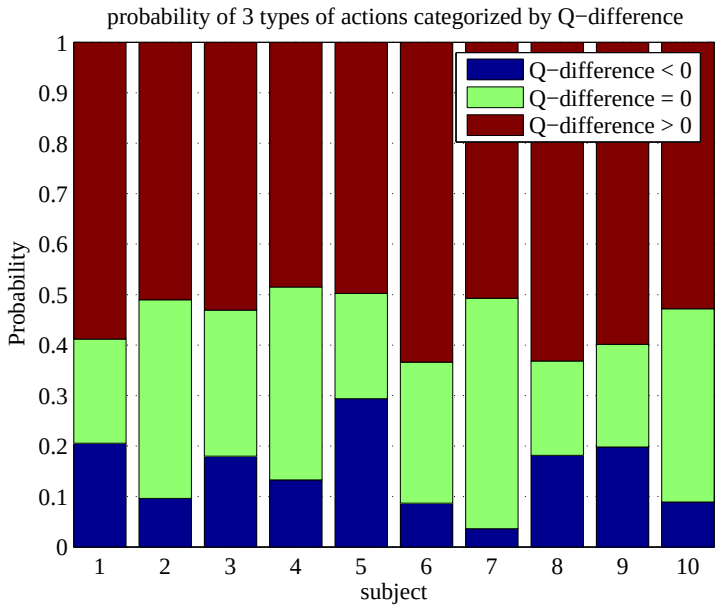


図 4.27　実験 1 の全タスクを通しての全行動における Q 値の差分によって分けられた行動確率

結果 1

図 4.27 より，被験者間に Q 値の差分ベースで分類した行動確率に有意な差が存在することが確認できる．特に，Q 値の差分が 0 より大きい行動確率よりも，0 以下である行動確率に相対的に大きな差が見られた．Q 値の差分が 0 未満の行動について，最も大きい被験者間の差として，被験者 5 が，全行動のうち約 30%が Q 値の差分が 0 未満の行動であるのに対して，被験者 7 はわずか約 4%であった．

考察 1

図 4.27 が示す行動確率は，Q 学習のアルゴリズムに従って，または反して行動した行動数の確率であるとも考えることができる．Q 学習アルゴリズムにおいて，エージェントは基本的には，ある状態における最も Q 値の高い行動を選択する．よって，仮にこの

greedy な行動規則が Q 学習エージェントにとっての「合理的」な行動決定パタンであるとすると[*2]，Q 値の差分が 0 より大きい行動は,「合理的」な行動，逆に Q 値の差分が 0 より小さい行動は「非合理的」な行動と呼ぶことができる．

また，被験者ごとのばらつきが大きい点については，Q 学習の観点から見つめると，それぞれが異なった行動決定に関する戦略を有しているという可能性が示唆される．

さらに，図 4.27 における「非合理的」な行動確率と図 4.18 に示した被験者間の学習結果の相関を見るために，各被験者が実験 1 の全 10 種類のタスクの遂行に要したステップ数と，それぞれの被験者が Q 値の差分が 0 未満の行動を選択した確率を，分布図によって表した (図 4.28)．

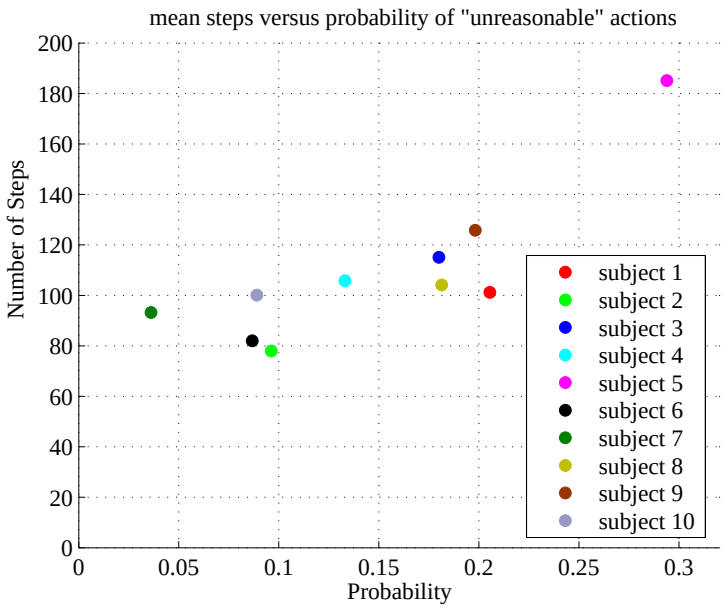


図 4.28　実験 1 における各被験者の学習結果と Q 値の差分が 0 未満の行動確率

結果 2

図 4.28 より,「非合理的」，つまり Q 値の差分が 0 未満の行動確率が高いほど，タスク遂行に要する平均ステップ数が増加する傾向が確認された．

[*2] 計算機プログラムの強化学習においても，エージェントは Q 値が相対的に小さい行動を選択することがある．これは最適化問題などにおいて局所解に陥ってしまうのを防ぐためで，エージェントは greedy な行動決定ルールと合わせて，確率的な行動決定を行なう仕組みを有している (2.4.4 項参照)．

考察 2

この結果より，実験 1 のタスクにおいて，また設定したパラメタにおいて，Q 学習エージェントとしての「非合理的」な行動を多く選択することは，タスク遂行に要するステップ数を増加させる，つまり学習の出来を悪化させる行動決定であるという自明な結果が確認された．

4.2.4 計算機の Q 学習プログラムとの比較

本項では，被験者と計算機の Q 学習エージェントとの行動決定の違いを探るための解析を行なう．ここで，人間である被験者と計算機プログラムである Q 学習エージェントには，以下のように大きく 2 つの相違点を挙げることができる．

- 環境のモデルに関する知識の正確さ
  被験者の場合，獲得したタスクの状態遷移に関する知識は完全に正確なものであるとは限らない．一方，計算機の Q 学習エージェントは被験者とは異なり，色による状態提示を受けるわけではなく，状態をそれぞれにユニークな数値として記憶しており，またそれらすべての状態におけるすべての行動の Q 値を正確に蓄積している．
- 行動や状態の持つ意味
  計算機プログラムには単に数値としてだけ捉えられている状態や行動について，被験者にとっては特別に認知的な意味が存在する可能性がある．本実験タスクの設計において，なるべく被験者にそのような意味づけが行なわれないように配慮した．しかし，被験者が行動の種類であるボタンの右，左や，状態を表している色に，なんらかの意味づけを行なっている可能性は否定できない．これは，被験者が実行する，複数の行動系列についても同様に考えられる．

以上のような相違点が挙げられる中で，被験者と計算機の強化学習エージェントとの行動決定の違いをマクロな視点から調査する．そのため，被験者に課したタスクと同一のタスクを計算機の Q 学習エージェントに課すシミュレーション実験を行ない，平均学習曲線，全行動の Q 値の差分の分布，また学習結果としての平均ステップ数について，被験者の結果との比較を行なう．

ここで，Q(λ) 学習のメタパラメタの設定は，逆温度 β=$\{5, 10, 20, 50\}$ の 4 種類を試行し，その他のパラメタについては被験者を解析した際の数値と同一とした．

シミュレーション結果 1：平均学習曲線

タスク a とタスク d の 2 種類のタスクについて，その平均学習曲線と各試行間の標準偏差を求めた (図 4.29，図 4.30)．なお，試行回数はそれぞれ 100 回である．

これらの図より，タスク a，タスク d のいずれのタスクにおいても，β の値の増加に伴って学習の収束が早まり，また試行ごとの分散が減少している様子が確認された．また，両者のタスク間の比較により，β が同一の値を持つときの学習曲線はほぼ同一であることが確認された．

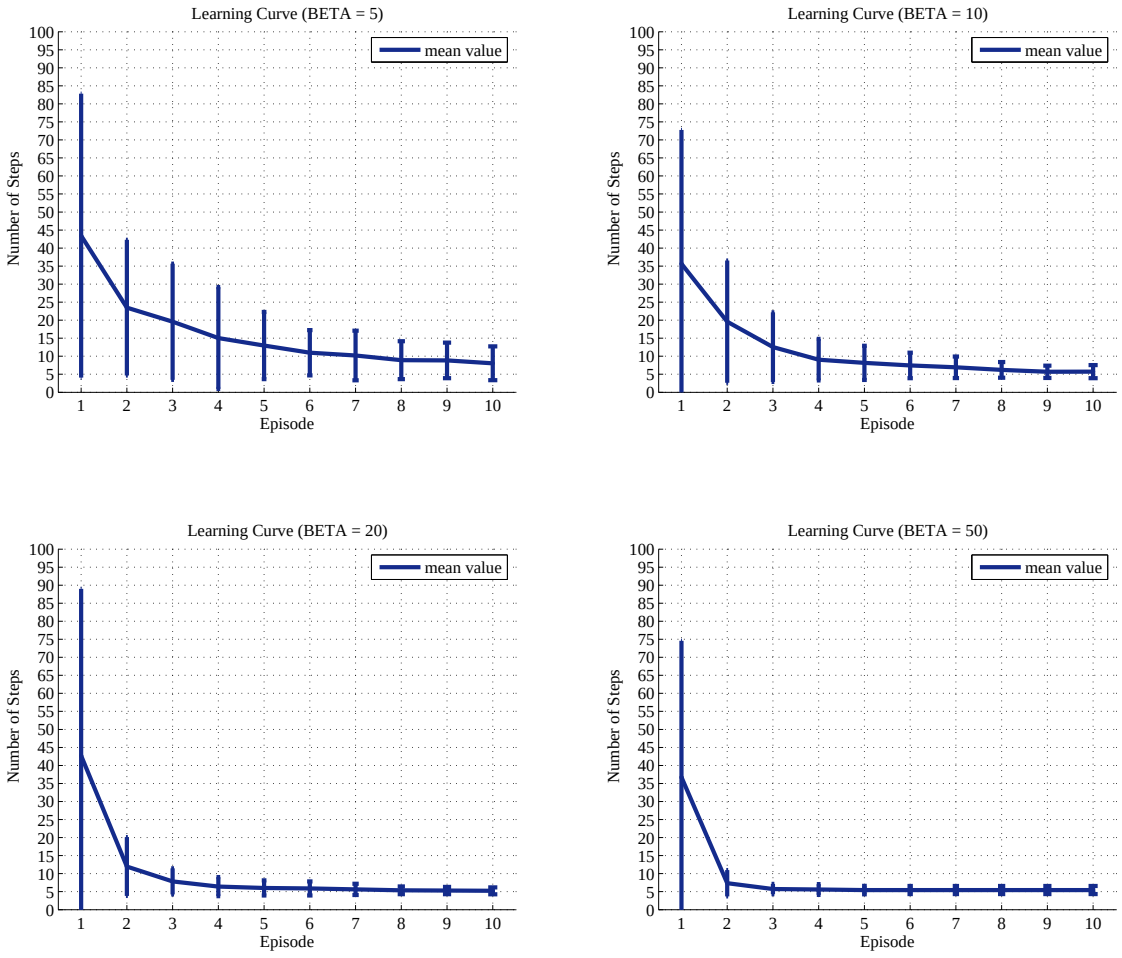


図 4.29　実験 1 タスク a の平均学習曲線 (左上:β= 5，右上:β= 10，左下:β= 20，右下:β= 50)

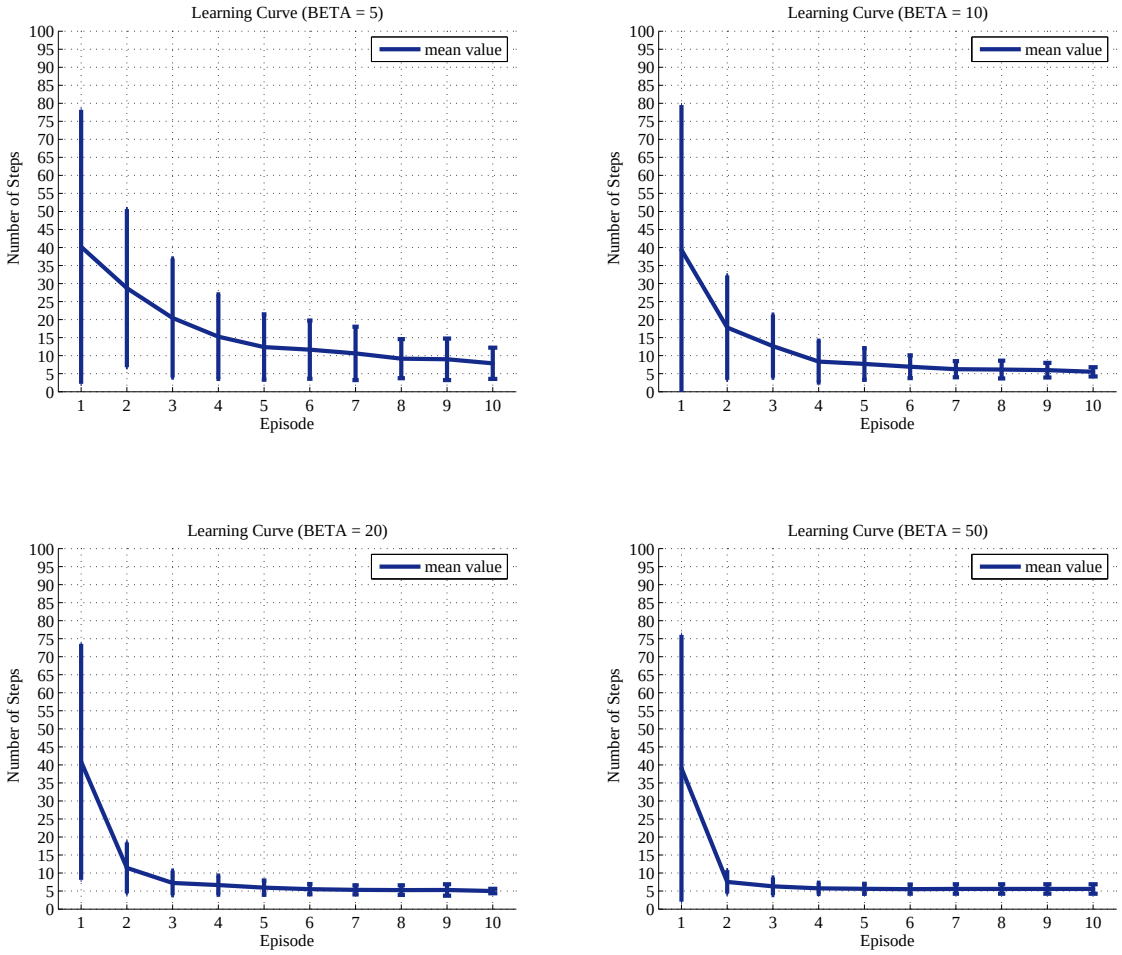


図 4.30　実験 1 タスク d の平均学習曲線 (左上:β= 5，右上:β= 10，左下:β= 20，右下:β= 50)

シミュレーション結果 2：全タスクを通しての全行動の Q 値の差分の内訳

次に，これらの全行動の Q 値の差分を図 4.27 に倣って分類した (図 4.31)．この結果より，β の値が高いほど，Q 値の差分が 0 未満の行動を取る確率が低いことが確認された．

シミュレーション結果 3：学習結果の差としての平均ステップ数

また，そのときの学習結果である，実験 1 の全 10 種類のタスクすべてを遂行したときに，1 つのタスクを遂行するのに要したステップ数の平均と，各タスク間におけるそれぞれのタスクの遂行に要したステップ数の標準偏差を示す (図 4.32)．

この結果から，β の値が高いほど，より少ないステップ数でタスクを遂行できたことが確認できる．また，図 4.31 とこの結果より，計算機の Q 学習エージェントにおいても，Q 値の差分が 0 未満である「非合理的」な行動確率が高いほど，タスク遂行に要するステップ数が増加するという傾向が確認された．

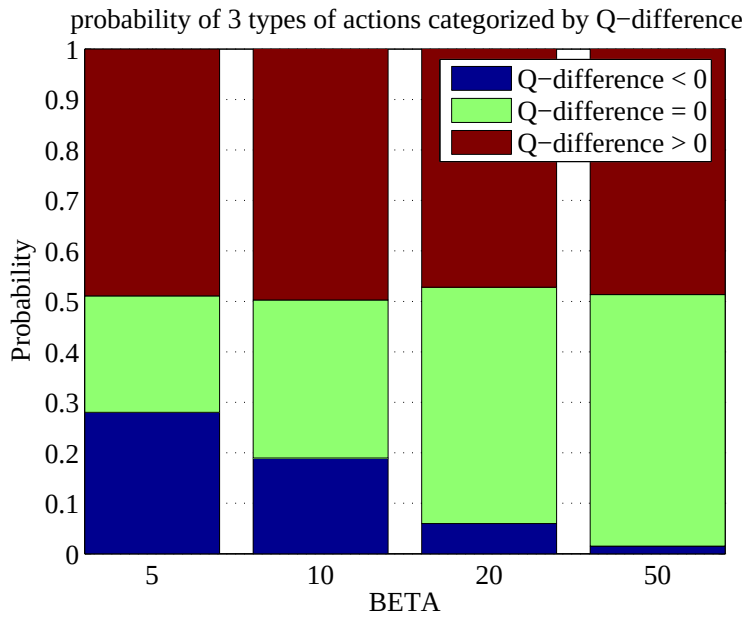


図 4.31　実験 1 の全タスクを通しての全行動における Q 値の差分によって分けられた行動確率

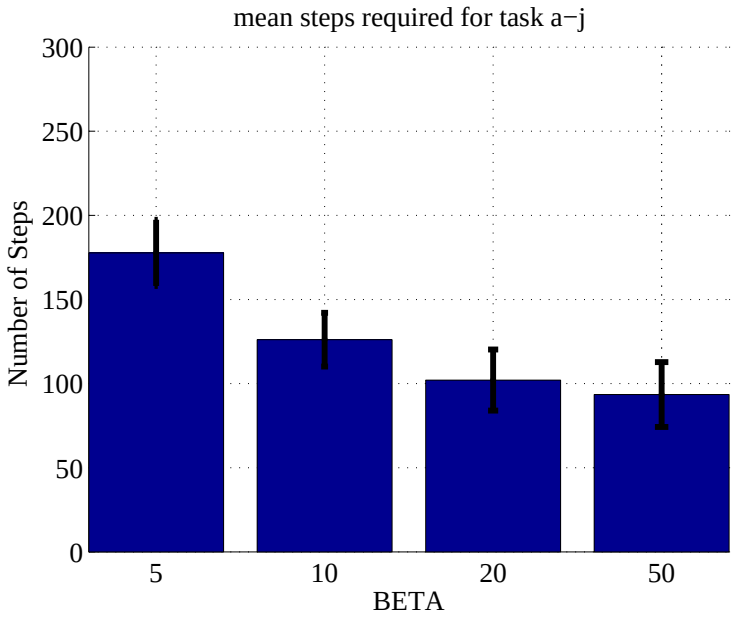


図 4.32　β の違いによる学習結果の差

シミュレーション結果 4：全タスクを通しての全行動の Q 値の差分の分布

さらに，実験 1 の全 10 種類のタスクにおける全行動について，その Q 値の差分の大きさによって分別された行動回数の分布を示す (図 4.33，図 4.34).

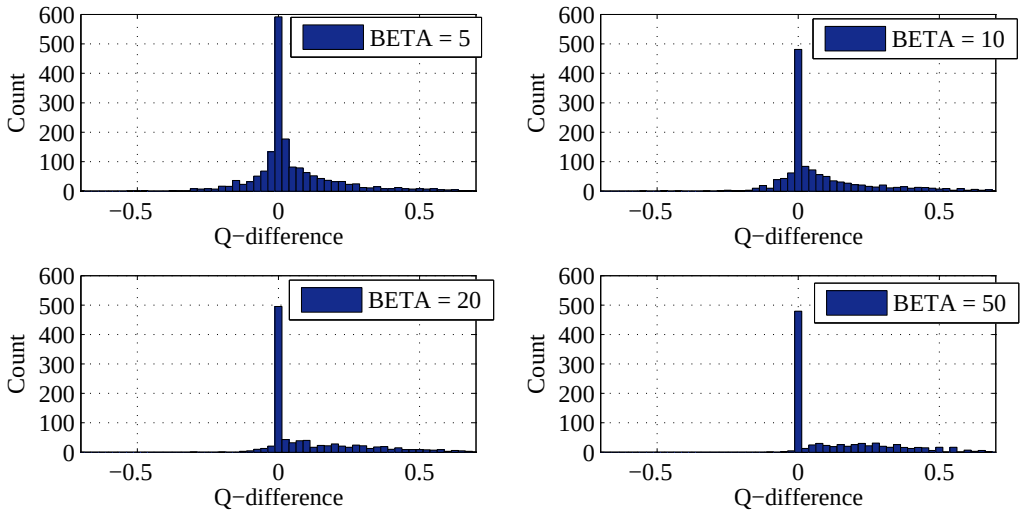


図 4.33　実験 1 の全タスクを通しての全行動の Q 値の差分

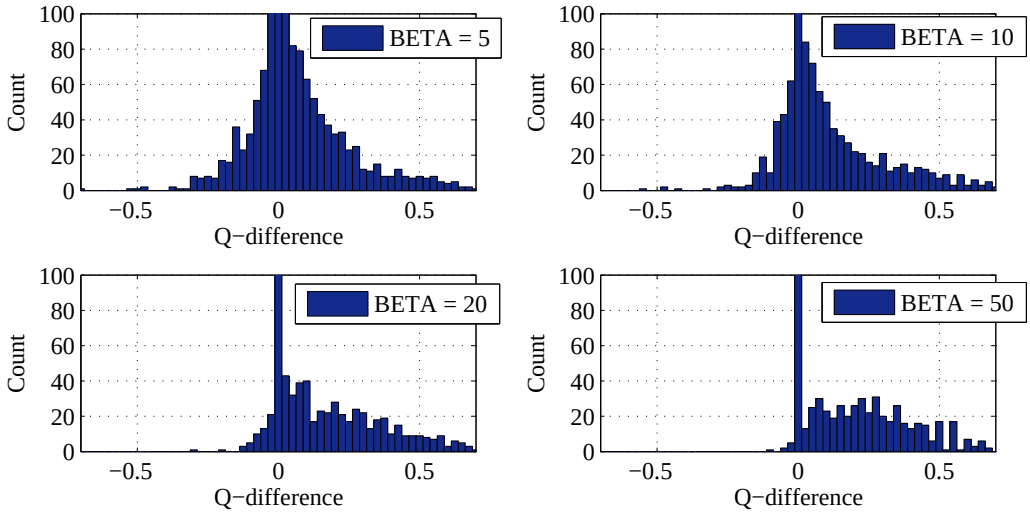


図 4.34　実験 1 の全タスクを通しての全行動の Q 値の差分 (図 4.33 の縦軸を拡大したもの)

被験者と計算機の Q 学習エージェントとの比較

以上の結果を用いて，被験者と計算機の Q 学習エージェントの行動決定の違いについて，主に Q 値の差分を用いたマクロな視点からの比較を行なった．具体的には，本項で得られたシミュレーション結果と，以前に示した被験者の行動から解析された以下のデータ群を用いて，両者の行動決定の様子を比較した．

- 平均学習曲線
- 全タスクを通しての全行動の Q 値の差分の内訳
- 全タスクを通しての全行動の Q 値の差分の分布
- 学習結果としての平均ステップ数

まず，被験者による平均学習曲線との比較を行なった．実験 1 のタスク a における被験者の平均学習曲線 (図 4.11) と，計算機の Q 学習エージェントの平均学習曲線 (図 4.29) を比較したところ，学習曲線の傾きが発散しないという点で，被験者と計算機の Q 学習エージェントの学習に共通点が確認された．一方，被験者にとって学習難易度が高いと示された，実験 1 のタスク d の平均学習曲線 (図 4.30) については，被験者の平均学習曲線 (図 4.13) と比較して，共通した傾向は見られなかった．

これらの結果により，被験者にとっては学習難易度に差があるタスク間においても，計算機の Q 学習エージェントにとっては，優位な差は存在しないことが確認された．

次に，Q 値の差分を用いた行動記録による，マクロな視点からの比較を行なった．図 4.27 と図 4.31 を比較すると，何人かの被験者における，Q 値の差分で分類した行動確率の内訳は，いずれかの β の値を持った計算機の Q 学習エージェントのものと非常に良く似ていることがわかる．例えば，被験者 5 は β=5 の Q 学習エージェントのものと，被験者 3 は β=10 の Q 学習エージェントのものと近い行動確率を有していることがわかる．これらの被験者，Q 学習エージェントについて，それぞれの Q 値の差分の分布の様子を比較すると (図 4.25，図 4.33 または図 4.26，図 4.34)，その分布の様子も近似していることが確認できる．さらに，学習結果としての，タスク遂行に要した平均ステップ数を比較すると (図 4.18，図 4.32)，それぞれの被験者とそれぞれの β を持った Q 学習エージェントについても，近いレベルでの一致がみられる．

これらの一致は，上記の被験者に特定されるものではなく，全体的な傾向として確認することが可能なものである．

以上の結果は，被験者の行動と，計算機の Q(λ) 学習エージェントの行動には，マクロな視点からの解析による一致がみられたということを示している．すなわち，本実験タスクの条件における被験者たちの行動決定は，計算機の Q(λ) 学習エージェントの探索行動

を行なう確率を決定するパラメタである β の値を変更することによって，マクロ的には説明可能であるということを示している．

eligibility trace の影響の調査

Q(λ) 学習における eligibility trace の影響を調査するため，Q(0) 学習によって同様の解析を行なった．このとき，メタパラメタの値として，α={0.3, 0.5, 0.7}，γ={0.5, 0.7, 0.9} の計 9 通りの組み合わせを試行した．以下に，その中で最も被験者の学習曲線に近い学習曲線を出力したパラメタである，α=0.7，γ=0.9 のときの Q(0) 学習エージェントの学習曲線を示す (図 4.35)．また，同一のパラメタにおける，Q(0) 学習エージェントと被験者の，Q 値の差分によって分別された行動選択確率をそれぞれ図 4.36，図 4.37 に示す．

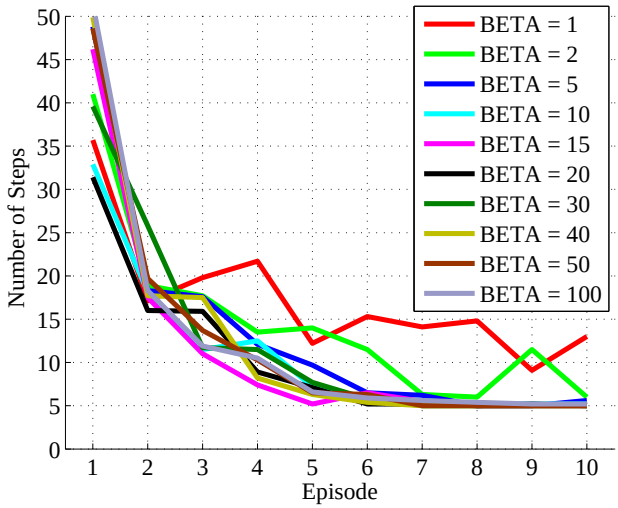


図 4.35　Q(0) 学習エージェントによる実験 1 の平均学習曲線 (α=0.7，γ=0.9)

図 4.35 より，試行した 9 通りのうち，もっとも被験者のものに近い学習曲線においても，被験者の平均学習曲線 (図 4.15) における被験者 2 や，図 4.29，図 4.30 における β=50 のエージェントのように，エピソードの前半から学習が収束するような結果は得られなかった．これは，Q(λ) 学習が Q(0) 学習と比較して，より少ないエピソード数で学習が可能であることを示していると同時に，被験者の学習は，Q(0) 学習エージェントと比較して Q(λ) 学習エージェントのものに近いことを示している．

また，図 4.36，図 4.37 より，Q(0) 学習エージェントは被験者と比較して，Q 値の差分が 0 より大きい行動選択確率が相対的に低く，双方で近似した行動選択確率の分布は得られなかった．この結果も，被験者の学習が Q(λ) 学習エージェントのものに近いことを支持する．

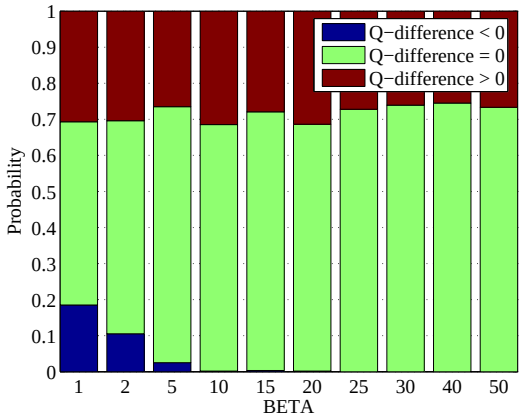


図 4.36　Q(0) 学習エージェントの行動選択確率

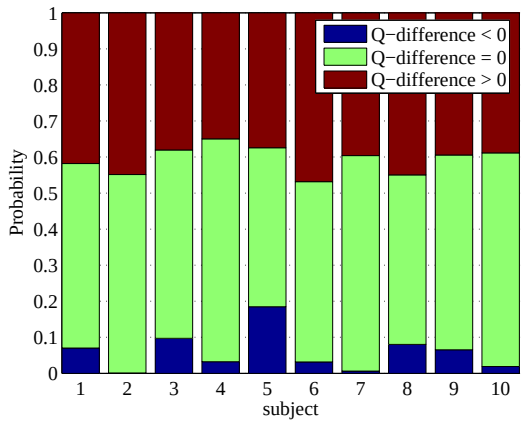


図 4.37　被験者の行動選択確率

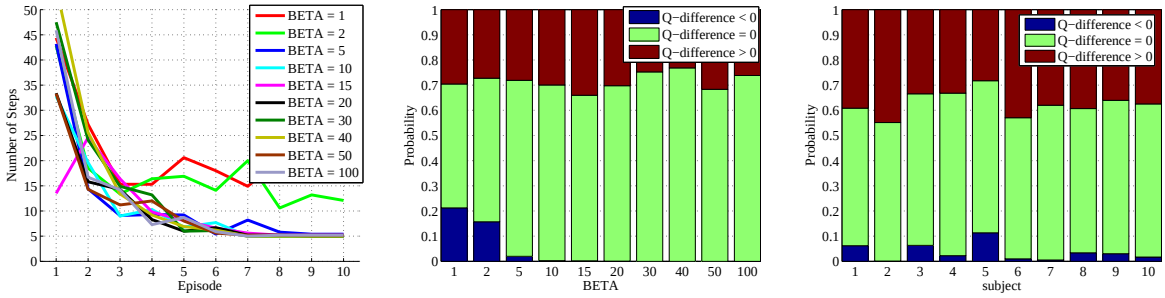
これらの結果より，本実験タスクにおける被験者の行動決定には，Q(λ) 学習における eligibility trace に相当する要素が関連していると考えることができる．すなわち，被験者は自分が過去に行動した系列になんらかの重み付けをしており，その記憶を用いて行動決定を行なっているという可能性が示された．

Q 学習以外のアルゴリズムによる解析

本研究における強化学習の枠組みから捉えた解析は，主に Q 学習アルゴリズムによって行なわれた．一方，先に述べたように，強化学習にはさまざまなアルゴリズムが存在する．それらにおける膨大なメタパラメタの組み合わせの中で，被験者の学習結果と一致する組み合わせを探すことは本研究の趣旨とは異なるが，ここでは，Sarsa と Actor-Critic および，それぞれに eligibility trace を実装した計 4 種類のアルゴリズムを用い，エージェントの平均学習曲線および Q 値 (または p 値) の差分によって分別された行動確率を求めた (図 4.38，図 4.39，図 4.40，図 4.41)．

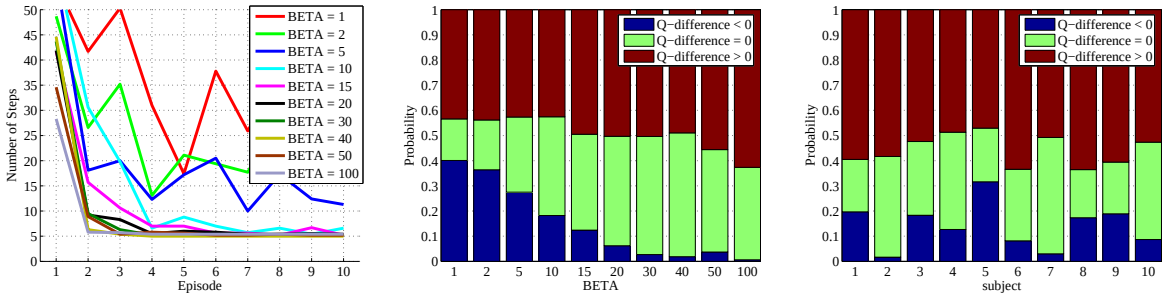
なお，各アルゴリズムにおけるメタパラメタは，先の解析における Q(0) 学習と Q(λ) 学習のものに倣って設定された 1 種類のみを試行した．ただし，Actor-Critic においては，critic，actor の学習率をそれぞれ α_1，α_2 とし，eligibility trace の減衰率をそれぞれ λ_1，λ_2 とした．

図 4.39 より，Sarsa(λ) エージェントの平均学習曲線において，β の値によってはエピソードの初期からその収束が確認され，被験者の平均学習曲線 (図 4.15) に近いものであることが示された．また，Q 値の差分によって分別された行動確率についても，被験者のシミュレーション結果と近い分布を持ったものであることが確認された．この近似は，4 種類のアルゴリズムのうちで最も高いレベルであり，もともと Q(λ) 学習と非常に良く似たアルゴリズムである Sarsa(λ) も，被験者の学習に近いものであることが示された．



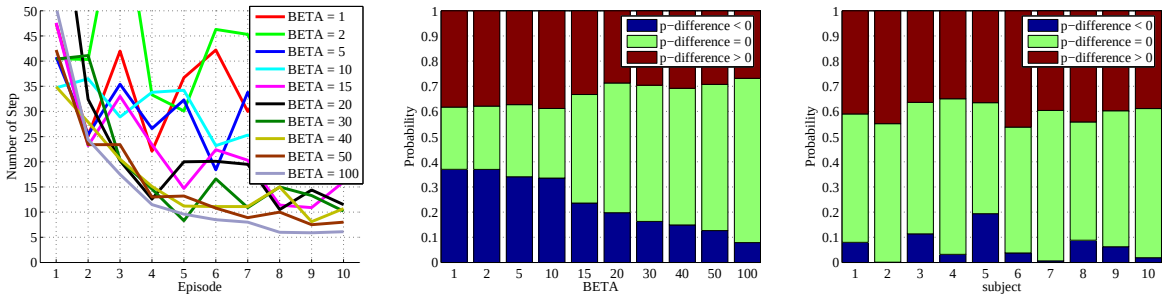
平均学習曲線　　エージェントの行動確率　　被験者の行動確率

図 4.38　Sarsa(0) による解析 (α=0.7, γ=0.9)



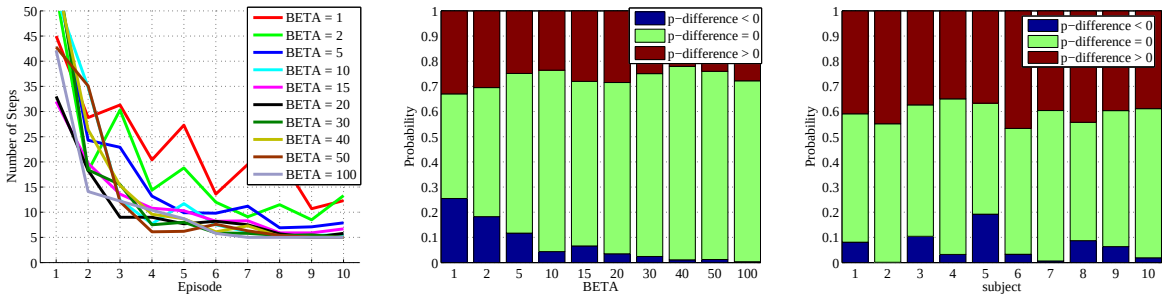
平均学習曲線　　エージェントの行動確率　　被験者の行動確率

図 4.39　Sarsa(λ) による解析 (α=0.1, γ=0.9, λ=0.9)



平均学習曲線　　エージェントの行動確率　　被験者の行動確率

図 4.40　Actor-Critic(0) による解析 (α_1=0.7, α_2=0.1, γ=0.9)



平均学習曲線　　エージェントの行動確率　　被験者の行動確率

図 4.41　Actor-Critic(λ) による解析 (α_1=0.1, α_2=0.1, γ=0.9, λ_1=0.9, λ_2=0.9)

4.3　人間の性質を考慮した解析

前節の結果より，被験者と Q(λ) 学習エージェントの学習はマクロ的には近似したものであることが示された．しかし，4.1.3 項より，被験者と Q 学習エージェントの間には，同一タスクにおいて，学習難易度に差が存在することが示された．

本節では，同一トポロジーのタスク間において，計算機の強化学習エージェントには見られなかった学習曲線や学習難易度の差が，被験者において認められた原因を探るため，被験者へのインタビューや強化学習の枠組みから離れた人間の性質に基づいて，被験者の行動決定に関する分析を行なう．

4.3.1　被験者へのインタビュー

実験 1 において，被験者はすべてのタスクを遂行した後に，実験に関する簡単なインタビューを受けた．これらの内容は，人間である被験者達の，強化学習の枠組みからは説明できない行動決定ルールに関する知見を得るための鍵となるかもしれない．以下にその代表的な内容を示す．

1.「行動決定が難しい，または簡単であると感じたタスクや状況はどのようなものであったか？」
   (a)「同じボタンを押し続けるだけでゴールできるタスクは簡単」
   (b)「似たような色の状態における行動決定が難しい」

   計算機プログラムの強化学習エージェントにおいては，状態や行動，または行動系列に特別な意味づけを行なっていないので，このような難易度の差は人間独特のものであると考えられる．

2.「どのような戦略でタスクを遂行していたか？」
   (a)「1 エピソード目でランダムに行動し，ゴール直前の状態の色とゴールへの行動を記憶しておき，次のエピソード以降はさらにその一つ前の状態-行動を一つずつ記憶していく」
   (b)「ゴールまでの正解系列がわからないとき，左または右のボタンを連続して押し，それだけでゴールに辿り着けるかどうかを確かめる」
   (c)「初期状態に戻ってきてしまったときに戦略を考え直す」

   (a) の意見は，大部分の被験者が回答していたものであり，状態-行動のペアでの記憶という意味において，Q 値による学習を行なう Q 学習アルゴリズムとの整合性が認められる．また，被験者がスタートからゴールへ辿ってきた状態遷移をすべて記憶

することができていない様子が示唆される．(b) は，行動系列についても意味づけが行なわれていることを示しており，被験者が探索行動を行なう前に，前もって特定の行動決定戦略を有している可能性を示している．(c) は，ゴールすることができないまま，初期状態という，認知的重みづけが相対的に強いと考えられる状態に戻ってきてしまったときに，行動決定の戦略を変更しようとしていることを示している．

3.「行動と状態遷移についてどのように記憶していたか？」
(a)「この色が出たらこのボタンを押す」
(b) (正解系列を発見してからは)「色ではなくリズム，またはマウスのクリック音」
(c)「色は言葉で，行動は手で記憶していた」

(a) は，極めて Q 学習アルゴリズム的な，状態-行動のペアで価値を判断して行動決定している例である．(b) や (c) は，状態や行動に認知的または身体的な意味づけが行なわれていたことを示している．また，大部分の被験者が，(a) と (b) の両方の意見を回答しており，それらの (別々であると思われる) 記憶がエピソードの進行に伴って合致していくという意見も得られた．これは，後ろ向きに獲得した状態-行動ペアに関する短期記憶と，前向きに獲得した正解系列に関する手続き型の記憶[*3]が融合しているという可能性を示している．

4.「どのようなときに探索行動を行なったか？
(a)「ゴールに要したステップ数が 7 以上のとき」

被験者は最短ステップ数が 5 であることを知っているので，より少ないステップ数でゴールしたいという希望の表れであると同時に，前回のエピソードにおける成績 (ステップ数) と未来のエピソードにおける成績の差分が被験者への「報酬」となっている可能性を示している．

5.「ミスであると思った行動決定はどのような理由によるものであったか？」
(a)「単純なボタンの押し間違い」
(b)「記憶があやふやであった」

これらの意見は，被験者が本実験タスクを実行する場合，記憶の欠如や運動ミスといった，計算機プログラムには存在しない要素が関係していることを示している．

6.「タスクの状態遷移図をどのようにイメージしていたか？」
(a)「二分木」
(b)「ワープの存在する 2D の迷路」

[*3] 運動や技能など，動作を繰り返し経験することによりその規則性を学習，獲得するもの．

4.3.2 応答時間

被験者間の行動決定に関する差異を調査するため，各行動決定に要した応答時間による分析を行なった．ここで，応答時間とは，状態がモニタに提示されてからボタンを押すまでの時間のことである．つまり，状態を観測してから行動を決定し，その行動を実行するまでの時間ということになる．

応答時間の概念は，一般的な強化学習のフレームワークには存在しない．しかし，本タスクにおける被験者の行動決定において，その判断の難しさを表す一つの指標となる可能性が考えられる．なぜなら私達は，通常の生活においてなんらかの行動決定を行なう際，その判断の難しさに比例して，行動決定にかかる時間が大きくなると考えられるからである．

一般的に，行動決定の判断基準となる要因や行動の選択肢が多ければ多いほど，またその選択肢間の重み付けの大きさが近ければ近いほど，行動決定に関する判断は難しくなると考えられる．また，その決定によって及ぼす影響の重大さも，判断の難しさに影響を及ぼすものと考えられる．

行動決定の種類の違いによる応答時間の増減

図 4.42 に，あるタスクにおける応答時間の増減の典型的な一例を示す．

結果 1

図 4.42 において，エピソードの 3 回目からタスクの終了まで，応答時間が一定して低い値が保たれていて，またそのばらつきも少ないことが確認できる．またその間は，すべて最短のステップ数である 5 ステップでゴールしていることが確認できる．

考察 1

この例において，被験者は，エピソードの 3 回目において最短の正解系列を発見し，タスクの終了まで同一の正解系列を選択しつづけている．つまり，この被験者は，前半のエピソードにおける探索行動によって獲得した知識に基づいて，エピソードの 3 回目以降は搾取行動のみを行なっていると捉えることができる．この結果は，搾取行動は探索行動と比較して，その行動決定に関する応答時間が小さいということを示している．

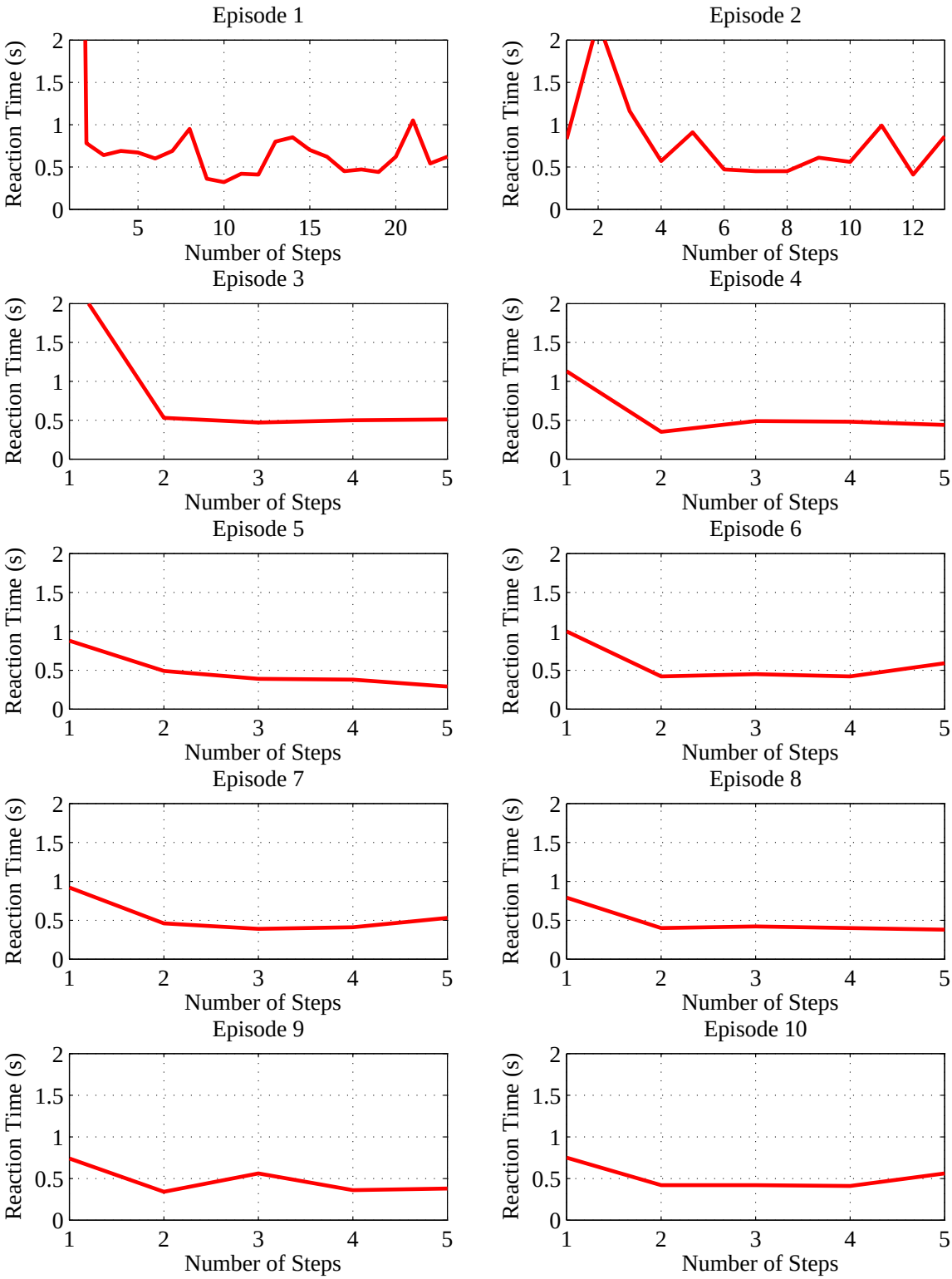


図 4.42　実験 1 タスク a における応答時間の遷移 (被験者 1)

また，実験 1 のタスク a におけるすべての被験者の応答時間の，エピソードごとの平均値の遷移の様子を示す (図 4.43).

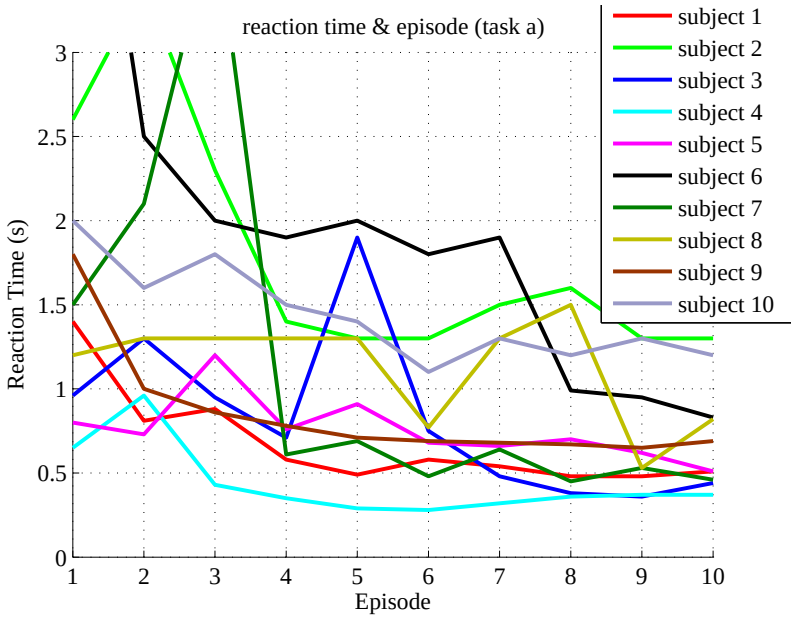


図 4.43　エピソードの増加に伴う応答時間の平均値の遷移の一例 (実験 1 タスク a)

結果 2

図 4.43 より，多くの被験者において，エピソードの増加に伴う応答時間の減少傾向が認められる．また，全体的には減少傾向にあるが，被験者 2 や 7 などについては，応答時間が一旦上昇した後に大きく下降している様子が確認できる．

考察 2

被験者全体の応答時間の減少傾向については，考察 1 より，エピソードの増加に伴って行動決定数に占める搾取行動の割合が高まったことが原因であると推察できる．また，被験者 2 と 7 については，前項のインタビューにより (4.3.1 2(a))，1 エピソード目における行動決定の戦略の一つとして，ゴールに辿り着くまでランダムな行動を行なうという戦略を取っている可能性が考えられる．なぜなら，これらの被験者は，応答時間の下降する直前のエピソードにおいて最短の正解系列を発見しており (図 4.1)，1 エピソード目では行動決定のための思考に大きな時間を要さない，つまり比較的応答時間の小さいランダムな行動決定をし，以降のエピソードでは時間をかけて慎重にゴールへの経路を探索し，最短の正解系列を発見した後は搾取行動を取り続けていると推察することが可能であるからである．

応答時間の分布

また，実験 1 の全タスクにおける，すべての行動の応答時間の分布について調査した (図 4.44).

結果

図 4.44 より，被験者間に応答時間の分布に，ばらつきが存在することが確認された．また，何人かの被験者においては，100〜200msec という非常に小さい応答時間の行動が観測された．

考察

被験者がモニタに表示される色によって状態を認識し，それに対応した行動を決定してマウスのボタンを押下するまでの時間として，100〜200msec という時間はあまりにも短すぎると考えられる．従って，このような非常に小さい応答時間を持つ行動は，最短正解系列を学習した後の搾取行動などであると推察される．

Q 値の差分と応答時間の相関

前節の結果を受けて，被験者が計算機プログラムの強化学習エージェントのように，マクロ的に見て Q 値に従った行動決定を行なっていると仮定すると，行動の合理性の目安となる Q 値の差分と応答時間との間に，何らかの相関が見つけられるかもしれない．そこで，実験 1 の全タスクにおけるすべての行動の Q 値の差分と応答時間を，分布図によって表した (図 4.45).

結果 1

図 4.45 より，被験者 4 を除くすべての被験者について，Q 値の差分が 0 のときを頂点とした分布の山が形成されていることが確認できる．

考察 1

この結果より，Q 値の差分の絶対値が大きい行動は，応答時間が相対的に小さい傾向が存在することが示された．先に述べたように，Q 値の差分が大きい行動は，ゴールに近い状態かつエピソードの後半における行動である可能性が高いので，行動決定について大きな時間を要さず，獲得した正解系列をひたすら繰り返すような搾取行動が，以上の傾向を形成していると考えることができる．

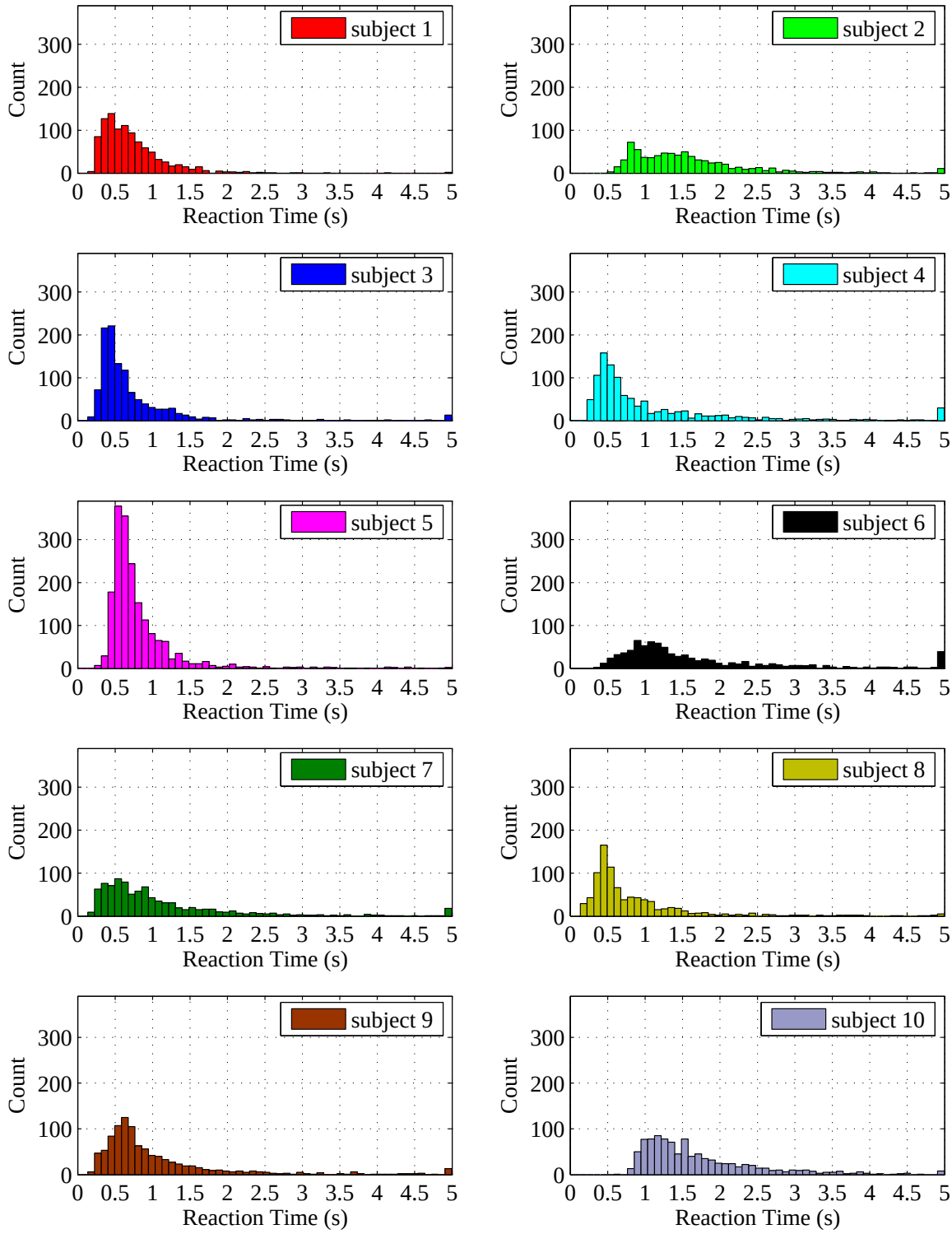


図 4.44　実験 1 の全タスクを通しての全行動の応答時間の分布

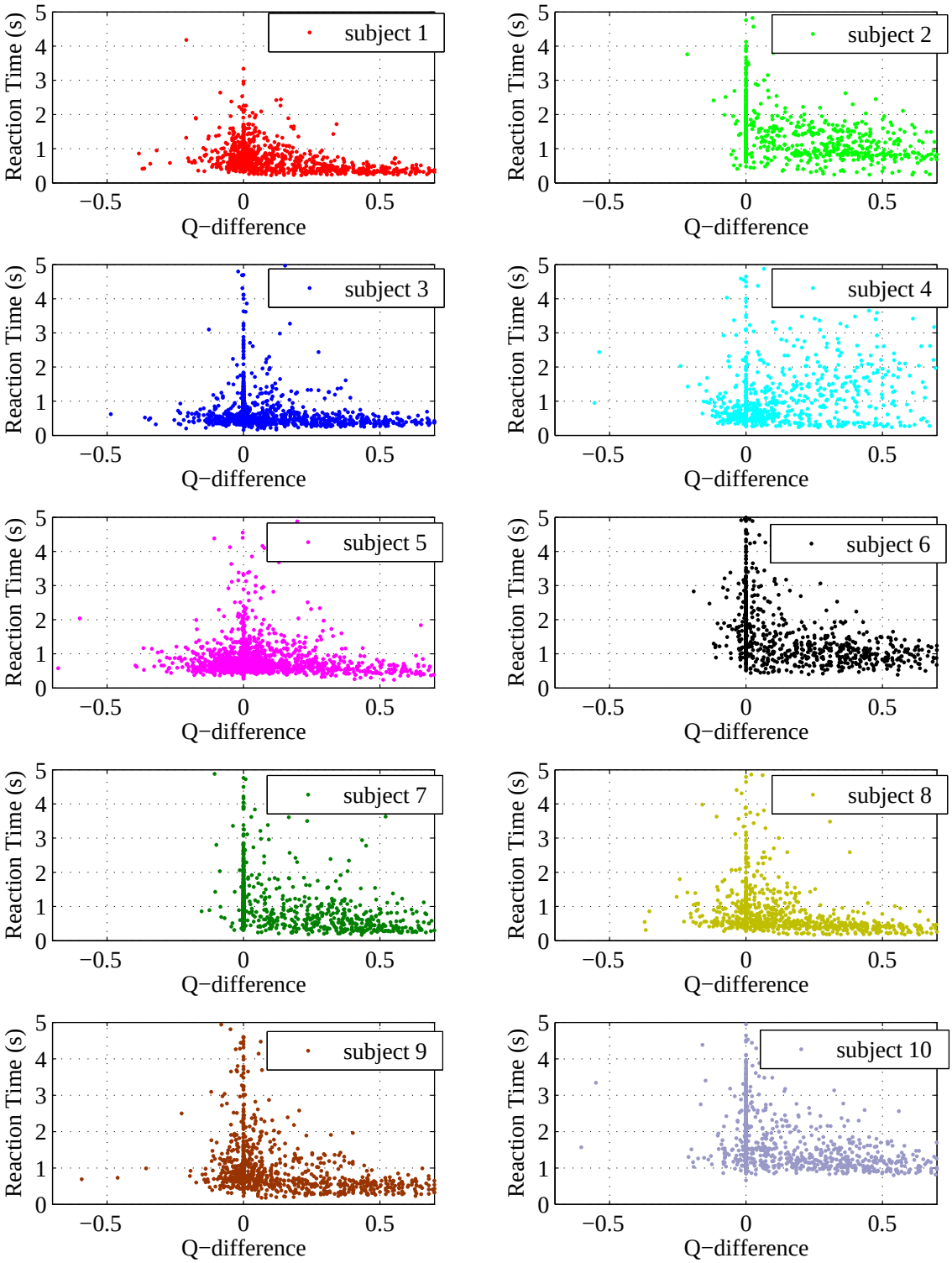


図 4.45　Q 値の差分と応答時間 (実験 1)

また，Q 値の差分と応答時間の関係をミクロに観察するために，図 4.46 に，実験 1 のタスク a における，Q 値の差分と応答時間のエピソードごとの動きを示す．

結果 2

図 4.46 より，エピソードの後半に進むほど，それぞれの行動における応答時間のばらつきが減少しており，またその Q 値の差分の大きさが増加していく様子が確認できる．

考察 2

この例において，被験者はエピソードの 5 回目に学習が完全に収束しており (図 4.1)，それ以降の搾取行動において，小さい応答時間で高い Q 値の差分を持つ行動を選択し続けていることが以上の結果より確認できる．これは，考察 1 で述べた Q 値の差分と応答時間の相関を支持するものである．

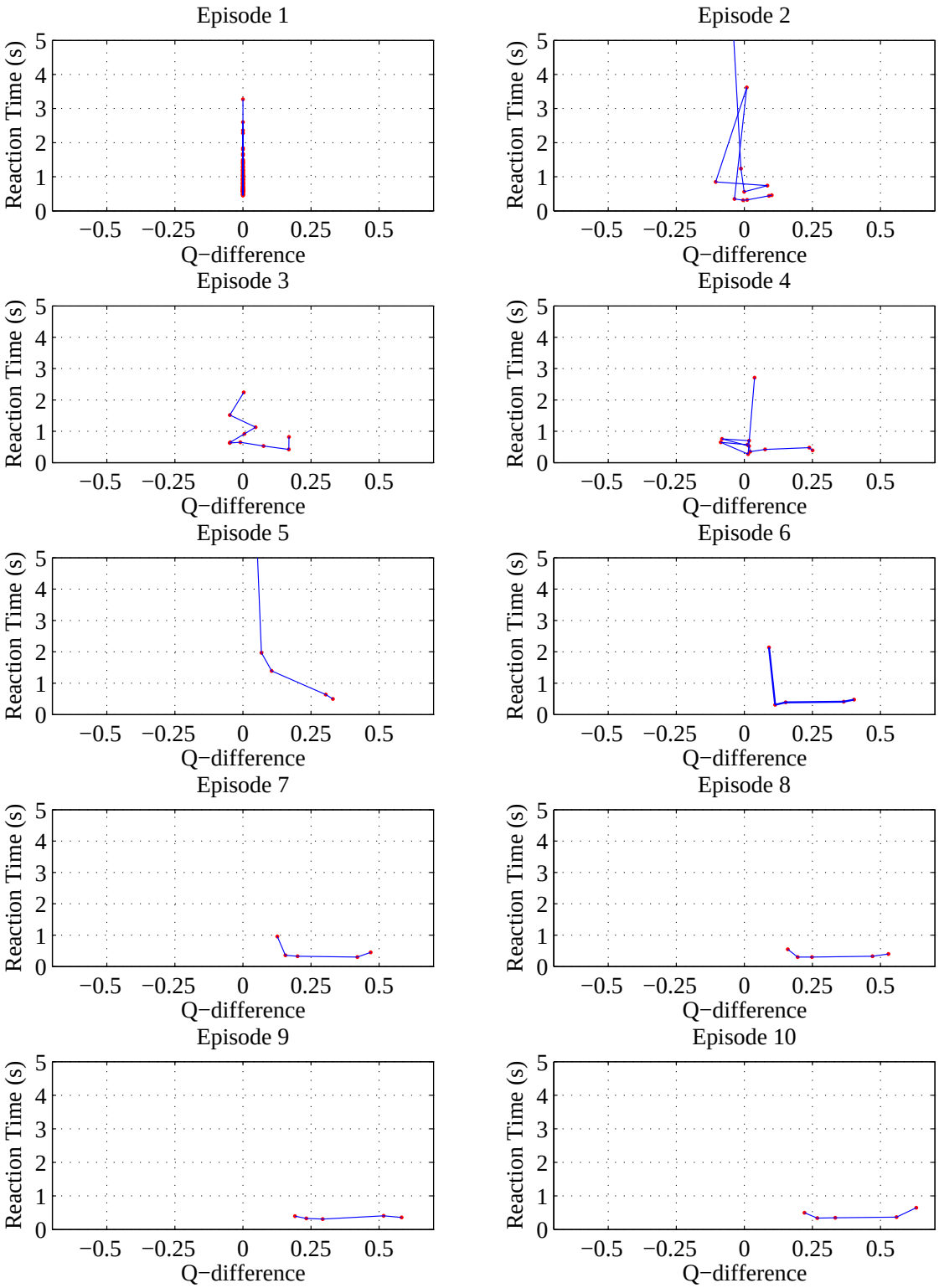


図 4.46　Q 値の差分と応答時間 (実験 1 タスク a) (被験者 3)

4.3.3 正解系列

4.1.3 項の分析結果より，同一のトポロジーであるタスク間に，学習難易度の相対的な差が存在することが示された．また，4.2.4 項の分析結果より，計算機の強化学習エージェントにはそのような学習難易度の差は存在しないことが確かめられた．一方，被験者へのインタビューより，同じ行動を選択し続ける (マウスの右または左ボタンを連続で押す) ことによってゴールできるタスクは，簡単であると感じたとの意見が得られた．

そこで，被験者にとってタスク間に学習難易度が存在する原因を探るため，タスクの持つ 2 種類の最短正解系列と，そのスイッチ回数に注目した分析を行なう．表 4.1 に，実験 1 のそれぞれのタスクが持つ 2 種類の最短正解系列と，10 エピソード目でいずれかの正解系列を選択した被験者の番号を示す．

結果 1

表 4.1 より，実験 1 の全 10 種類のタスクの中で，スイッチ回数の少ない最短正解系列を選択した被験者の数がより多かったタスクは 6 種類であり，そうでないタスクは 4 種類であった．また，いくつかのタスクにおいて，いずれかの最短正解系列を選択した被験者の数に大きな偏りが見られた．例えば，タスク e において，被験者 2 以外の 9 人の被験者すべてが，スイッチ回数が 1 である最短正解系列を実行している．同様に，タスク f やタスク g においても大きな偏りが確認された．

考察 1

以上の結果より，2 種類の最短正解系列のうち，選択した被験者数とスイッチ回数には，強い相関関係はみられなかった．しかし，以下に挙げるいくつかの特別な例においては，その相関がスイッチ回数を用いて説明できる．

タスク e やタスク j は，スイッチ回数が 0 と 3 という，10 種類のタスクの中では最も離れたスイッチ回数を持った最短正解系列を持つタスクである．このようなタスクにおいて，スイッチ回数が少ない最短正解系列を選んだ被験者が多数を占めた原因については，被験者の行動決定に関する戦略や，正解系列自体の覚えやすさといった理由を挙げて考えることができる．4.3.1 項の被験者へのインタビューによると，探索の際に，同じ行動 (ボタン) を選択し続けるという行動戦略を持つ被験者は多く，スイッチ回数が 0 であるような単純な最短正解系列は，このような戦略を持つ被験者にとっては，エピソードの 1 回目から発見できる可能性が存在する．また，そのような戦略を持たない被験者にとっても，スイッチ回数が 0 である正解系列は，最も覚えやすい行動系列であると考えられるので，記憶への定着も早く，搾取行動をより確実に行なうことができると考えられる．

	最短正解系列	スイッチ回数	選択した被験者
タスク a	L-L-L-L-R	1	1, 2, 3, 4, 5, 7
	R-R-L-L-R	2	9, 10
タスク b	R-R-L-L-L	1	2, 4, 8, 9
	L-L-R-L-L	2	1, 3, 5, 6, 7
タスク c	L-R-R-R-R	1	2, 3, 4, 6, 7, 8
	R-L-L-R-R	2	1, 5, 9, 10
タスク d	L-R-L-R-L	4	1, 2, 3, 5, 7, 9
	R-L-R-R-L	3	6, 8
タスク e	L-L-L-L-L	0	1, 3, 4, 5, 6, 7, 8, 9, 10
	R-L-R-L-L	3	2
タスク f	L-R-R-R-L	2	1, 2, 3, 4, 5, 7, 8, 10
	R-L-L-R-L	3	6, 9
タスク g	R-L-L-R-R	2	1
	L-L-R-R-R	1	2, 3, 4, 5, 6, 7, 8, 9, 10
タスク h	L-R-R-L-L	2	1, 2, 4, 7, 8
	R-R-R-L-L	1	3, 6, 9, 10
タスク i	L-R-L-L-R	3	2, 3, 4, 7, 8, 10
	R-L-R-L-R	4	1, 6, 9
タスク j	L-R-L-R-R	3	2, 5, 7
	R-R-R-R-R	0	1, 3, 4, 6, 8, 9

表 4.1　各タスクの最短正解系列と 10 エピソード目でそれらの正解系列を実行した被験者 (実験 1)

タスク f やタスク g における偏りについては，それぞれの最短正解系列のスイッチ回数が他のタスクと比較して特別に離れているというわけではなく，このような大きな偏りが見られた必然的理由は，スイッチ回数だけから説明することはできない．

また，4.1.1 項において，タスク d やタスク i が他のタスクと比較して，相対的に学習難易度が高いと示された点については，最短正解系列スイッチ回数の違いからの説明が可能であると思われる．なぜなら，これらのタスクは，どちらの最短正解系列とも，他のタスクにおける最短正解系列と比較して最も大きいスイッチ回数 (3，4) を持つ特殊なタスクであったからである．

以上の結果は，計算機の強化学習エージェントでは見られない傾向であり，人間独自の行動決定に影響を及ぼす要素であるということができる．

同様に，実験 2 および 3 について，最短正解系列とそのスイッチ回数に関する分析を行なった (表 4.2，表 4.3).

	最短正解系列	スイッチ回数	選択した被験者
タスク a	L-L-R-L-R-L-R	5	13, 16, 18
	R-R-L-R-R-L-R	4	11, 14, 19
タスク b	R-R-L-R-L-R-L	5	11, 12, 13, 14, 16, 20
	L-L-R-L-L-R-L	4	15, 17, 18, 19
タスク c	L-L-L-R-L-L-R	3	11, 13, 14, 18, 19, 20
	R-R-R-R-L-L-R	2	12, 15, 16, 17
タスク d	R-R-R-L-R-R-L	3	11, 13, 14, 16, 17
	L-L-L-L-R-R-L	2	12, 15, 18
タスク e	L-R-L-R-L-R-L	6	11, 13, 14, 16
	R-L-R-R-L-R-L	5	12, 18, 20
タスク f	R-L-R-L-R-L-R	6	19
	L-L-R-L-R-L-R	5	11, 12, 13, 16, 18, 20
タスク g	L-R-R-L-R-R-L	4	11
	R-R-R-R-R-R-L	1	12, 13, 14, 15, 16, 19, 20
タスク h	R-L-L-R-L-L-R	4	11, 16
	L-L-L-L-L-L-R	1	12, 13, 14, 15, 18, 19, 20
タスク i	L-R-R-R-R-R-L	2	11, 12, 13, 14, 16, 18
	R-L-R-L-R-R-L	5	20
タスク j	R-L-L-L-L-L-R	2	11, 12, 13, 14, 16, 18, 20
	L-R-L-R-L-L-R	5	

表 4.2　各タスクの最短正解系列と 10 エピソード目でそれらの正解系列を実行した被験者 (実験 2)

結果 2

表 4.2 より，実験 2 において，被験者は全体的にスイッチ回数の少ない最短正解系列を実行していることが確認された．特に，タスク g，h，i，j のような，スイッチ回数が大きく離れた最短正解系列を持つタスクにおいては，その傾向が顕著に表れていることが確認された.

また表 4.3 より，実験 3 においても，タスク c やタスク d のような，スイッチ回数が大きく離れた最短正解系列を持つタスクにおいては，同様の傾向が確認された.

	最短正解系列	スイッチ回数	選択した被験者
タスク a	L-R-R-R-L-L-L-R-R	3	21, 25, 29
	R-R-R-R-R-L-L-R-R	2	23, 28
タスク b	R-L-L-L-R-R-R-L-L	3	26
	L-L-L-L-L-R-R-L-L	2	23, 25, 28, 30
タスク c	L-R-L-R-L-L-R-R-L	6	
	R-R-R-R-R-R-R-R-L	1	21, 23, 25, 26, 27, 28, 29, 30
タスク d	R-L-R-L-R-R-L-L-R	6	21, 22
	L-L-L-L-L-L-L-L-R	1	23, 25, 26, 27, 28, 30
タスク e	L-L-L-L-L-R-R-R-R	1	21, 22, 23, 26, 28, 30
	R-R-R-R-L-L-R-R-R	2	25, 27, 29
タスク f	R-R-R-R-R-L-L-L-L	1	22, 25, 27, 29
	L-L-L-L-R-R-L-L-L	2	21, 26
タスク g	L-L-R-L-R-L-L-R-L	6	27, 30
	R-R-L-R-L-R-L-R-L	7	22, 24
タスク h	R-R-L-R-L-R-R-L-R	6	21, 22, 29, 30
	L-L-R-L-R-L-R-L-R	7	24, 28

表 4.3　各タスクの最短正解系列と 10 エピソード目でそれらの正解系列を実行した被験者 (実験 3)

考察 2

以上の結果より，最短正解系列がより長いタスクである実験 2 および 3 においても，実験 1 と同様に，スイッチ回数が大きく離れた最短正解系列を持つタスクにおいて，被験者はスイッチ回数が少ない最短正解系列を選択する傾向が確認された．

また，表 4.2 より，実験 2 において，タスク e やタスク f は他のタスクと比較して，もっとも大きいスイッチ回数の最短正解系列を持つタスクであり，実験 1 の考察結果を踏まえると，学習難易度が相対的に高いと予想される．しかし，図 4.22 より，タスク e やタスク f の学習難易度が他のタスクと比較して特別に高いという傾向は見られない．同様に，表 4.3 より，実験 3 において，タスク g やタスク h はもっとも大きいスイッチ回数の最短正解系列を持つタスクであるが，図 4.23 より，それらのタスクの学習難易度が特別高いという傾向は確認できない．これらの結果は，本研究で行なった 3 種類の実験において，最短正解系列が長いタスク，すなわち，記憶すべき行動系列が多いタスクでは，そのスイッチ回数がタスクの学習難易度に与える影響が小さいことを示している．

4.3.4　被験者固有の行動戦略

前項の正解系列による分析や 4.3.1 項のインタビューから，被験者ごとに前もった特有の行動決定における戦略が存在する可能性が示された．そこで，本項では，被験者がタスクを開始して 1 エピソード目の最初に選択した行動の偏りを調査した (図 4.47)．

具体的には，実験 1 の全 10 種類のタスクにおいて，被験者の 1 エピソード目の最初の行動が左ボタンであった確率を求めた．また，その行動と最終的に選択される正解系列の間に相関関係が見られるかどうかを調査した (図 4.48)．図 4.48 は，実験 1 の全 10 種類のタスクにおいて，被験者の 1 エピソード目の最初の行動が左ボタンであった確率 (横軸) と，最終的に選択された正解系列の最初の行動が左ボタンから始まるものであった確率 (縦軸) を分布図によって表したものである．

結果

図 4.47 より，被験者 1 と被験者 8 においては，すべてのタスクにおける最初の行動が左ボタンを押す行動であったことが確認できる．また，被験者 10 においては，それらがすべて右ボタンであったことが確認できる．

また，図 4.48 より，被験者が最初に選択した行動と，最終的に選択するに至った正解系列の間に，相関関係は見られなかった (相関係数 0.36)．

考察

以上の結果より，タスク開始後の最初の行動を，常に前もって決定していると考えられる被験者が存在することが示された．また，被験者が最初に選択した行動と，最終的に選択された行動系列の間に相関は存在せず，被験者が探索行動を行なっていくうちに，最初に選択した行動とは異なる行動から始まる正解系列を発見している可能性が示された．

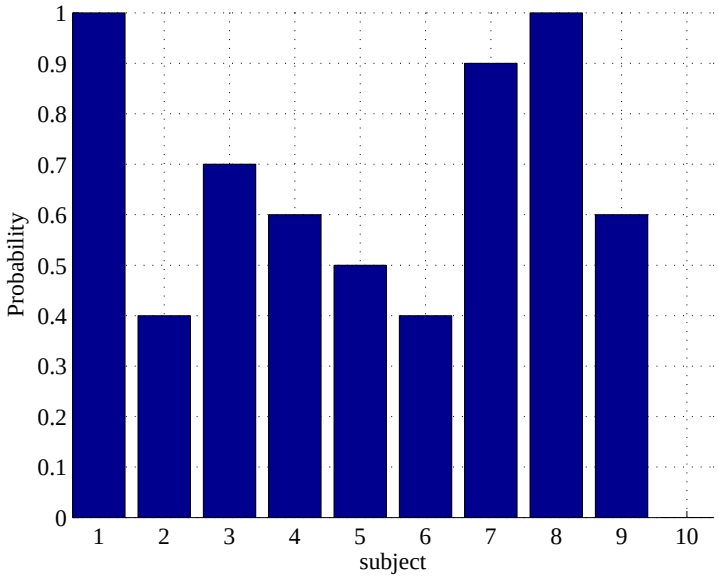


図 4.47　タスクの最初に左ボタンを押す確率 (実験 1)

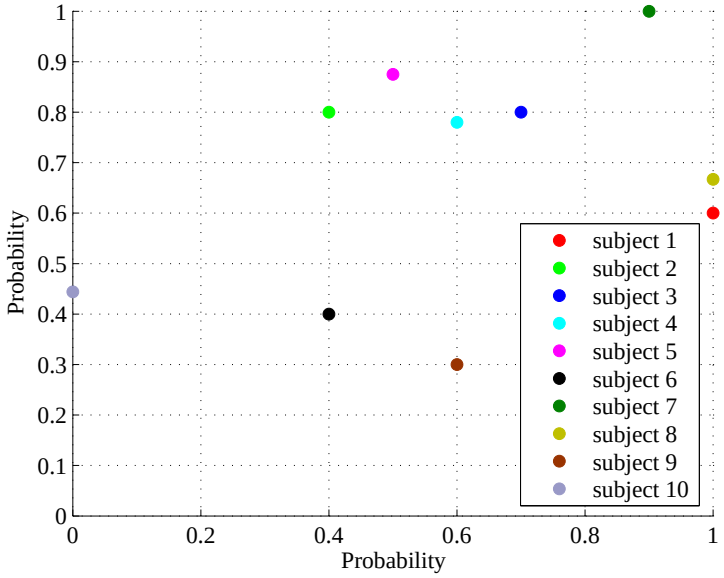


図 4.48　タスクの最初に左ボタンを押す確率 (横軸) と最終的に左ボタンから始まる正解系列を選択する確率 (縦軸) の関係 (実験 1)

4.3.5　プライミング効果

計算機プログラムの強化学習エージェントにとっては全く影響はないが，被験者は，実行するタスクの順序によって，前試行に取った行動が次試行の行動に何らかの影響 (プライミング効果) を受ける可能性が存在することが考えられる．これまでの分析結果から，本実験タスクにおいて被験者は「何色だからどちらのボタンを押す (ことが価値の高い行動である)」というように，状態-行動をペアで記憶している可能性がある．

そこで，最も強く記憶に残っていると考えられる，直前のタスクにおける最後の 5 ステップの状態-行動ペアの記憶を，次回のタスクにおいて模倣することをプライミング効果と定義し，その有無に関して調査を行なった．具体的には，実験 1 において，直前のタスクにおける 10 エピソード目の最後の 5 回の行動を，最初の行動，2〜4 回目の行動，最後の行動の 3 種類に分類した．そして，直前のタスクにおいて経験したそれらの状態 (色)-行動 (左右どちらのボタンを押したか) を，次回のタスクに初めて訪れた状態において模倣した確率を求めた (図 4.49)．

結果

図 4.49 より，直前のタスクの模倣行動を行なった確率に，被験者間の差が認められる．また，分類した行動間において，被験者ごとの相関が若干見られる．最も高い模倣行動の確率として，被験者 6，被験者 7 の最初の行動における，90%近い確率が認められる．ただし，全体的に観察すると，多くの被験者，また分類した行動において，その確率は 50%から大きく外れないものであることが確認された．

考察

被験者 6，被験者 7 の 90%に近い値は，本稿で定義するプライミング効果による影響である可能性を否定できない大きな値である．しかし，図 4.47 より，被験者 7 はタスクの最初に押すボタンが左ボタンであるような行動戦略を持っているとも考えられ，単純に直前のタスクの知識に引きずられた模倣行動であるとは断定できない．以上より，プライミング効果による影響が正または負の形で強く存在したと断定できる例は発見されなかった．

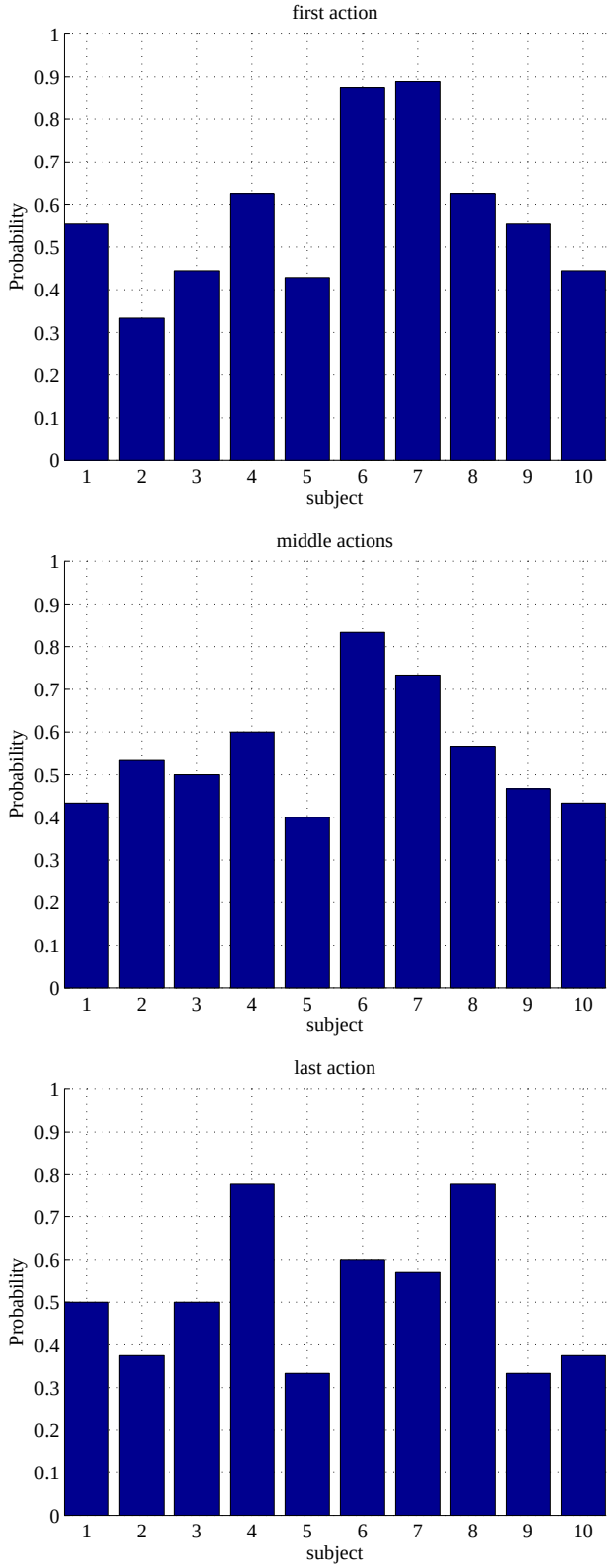


図 4.49　直前のタスクのゴール行動を反復した確率 (実験 1)
(上段:最初の行動，中段:2〜4 回目の行動，下段:最後の行動)

第 5 章

結論

本研究では，強化学習の枠組みから人間の行動決定を明らかにすべく，強化学習型タスクを設計し，被験者を強化学習エージェントと見立て，その行動決定に関する分析を行なった．その際，被験者の行動決定を，より広範な状態においてミクロな視点から観察するために，Q(λ) 学習を用いた解析を行なった．同時に，一つ一つの行動の合理性を定量的に表すための指標として，行動間の Q 値の差分を取るという方法を用いた．

このようにして得られた被験者の分析結果と，計算機プログラムの強化学習エージェントによるシミュレーション結果との比較を行なった．さらに，強化学習の枠組みから外れ，被験者の行動決定について，人間の認知的な性質を考慮した分析を行なった．

その結果，被験者の行動決定は，計算機の強化学習エージェントにおける，探索と搾取の行動確率を決定する 1 種類のパラメタのみを変化させることにより，マクロ的には説明することができた．また，被験者の行動決定は，Q(0) 学習と比較して Q(λ) 学習に近いものであり，被験者の学習に，eligibility trace に相当する記憶のメカニズムが関連していることが示された．

一方，個々のタスクを個別に見ると，計算機には簡単に学習可能であるが，人間にとっては難しいタスクが存在することが確認された．また，その原因としては，行動系列の覚えやすさの違いや，被験者独自の前もった行動戦略が挙げられることが示された．これにより，本実験タスクにおいて，人間の記憶のメカニズムなどの要因が学習結果に影響を及ぼしていることが示唆され，被験者の行動決定は，強化学習の枠組みからだけでは捉えきれないものであったことが示された．

本研究の背景としては，Schultz らをはじめとした，一連の脳の強化学習仮説に関する研究がある．本研究の出発点は，それらの仮説を受け，計算機プログラムで実行されるような強化学習問題を人間に課したら，一体どのような振る舞いをするのかという単純な好奇心であった．ただし，本研究で示された，被験者の行動決定がマクロ的に見て強化学習の枠組みから説明可能であったという事実は，脳の強化学習仮説を直接サポートするもの

ではないと考える．人間の行動や情動は単純なものではなく，それらのアウトプットと，実際の脳の内部モデルの相関を，本研究のような認知実験レベルで対応付けることは難しい．

本研究の趣旨とは異なるが，展望としては，Q 学習以外のさまざまな強化学習アルゴリズムを用いて被験者の行動のパラメタフィッティングを行ない，人間の行動決定に最も近いアルゴリズムやパラメタを調査することによって，より細かいレベルで人間の行動決定を議論するような研究が考えられる．

謝辞

本研究を進めるにあたり，ご指導，ご助言下さった阪口豊助教授に心から感謝の意を表します．同様に，出澤正徳教授，石田文彦助手，島井博行助手に心から感謝の意を表します．ならびに，ゼミ発表などでご意見や激励を下さったヒューマンインターフェース学講座の学生，研究生，修了生の皆様に心から感謝の意を表します．

参考文献

[1] 鮫島, 銅谷: “強化学習と大脳基底核”, バイオメカニズム学会誌, Vol.25, No.4, pp. 167-171, 2001.

[2] Sutton, R.S., Barto, A.G.: “Reinforcement learning: An introduction, a bradford book”, MIT Press, 1998. (邦訳 三上, 皆川: “強化学習”, 森北出版, 2000.).

[3] Skinner, B.F.: “Operant behavior”, American Psychologist, Vol.18, pp. 503-515, 1963.

[4] Barto, A.G.: “Adaptive critics and the basal ganglia”, Models of Information Processing in the Basal Ganglia, pp. 215-232, MIT Press, 1994.

[5] 彦坂, 山鳥, 河村: “眼と精神-彦坂典秀の課外授業”, 神経心理学コレクション, 医学書院, 2003.

[6] Schultz, W., Dayan, P., Montague,P.R.: “A neural substrate of prediction and reward”, Science, Vol.275, pp. 1593-1599, 1997.

[7] 設楽: “報酬の期待とモチベーションの脳内表現 前部帯状皮質と腹側線条体の神経活動”, 神経情報科学サマースクール NISS2002 講義録, 日本神経回路学会誌, Vol.10, No.2, pp. 84-89, 2003.

[8] Tanaka, C.S., Doya, K., Okada, G., Ueda, K., Okamoto, Y., Yamawaki, S. : “Prediction of immediate and future rewards differentially recruits cortico-basal ganglia loops”, Nature Neuroscience Vol.7, No.8, pp. 887-893, 2004.

[9] Suri, R.E., Schultz, W.: “A neural network model with dopamine-like reinforcement signal that learns a spatial delayed response task”, Neuroscience, Vol.91, No.3, pp. 871-890, 1999.

[10] Doya, K.: “Complementary roles of basal ganglia and cerebellum in learning and motor control”, Current Opinion in Neurobiology, Vol.10, No.6, pp. 732-739, 2000.

[11] 銅谷: “計算神経科学における強化学習「神経修飾物質系のメタ学習仮説」”, 人工知能学会全国大会第 16 回論文集, 2002.

[12] Singh, S.P., Sutton, R.S.: “Reinforcement learning with replacing eligibility

traces", Machine Learning, Vol.22, No.1-3, pp. 123-158, 1996.

[13] 藤崎: "強化学習型情報処理における人間の行動決定について", 情報処理学会研究報告「知能と複雑系」, Vol.123, No.24, pp. 135-140, 2001.

[14] 銅谷: "強化学習とメタ学習の脳内機構-大脳基底核と神経修飾物質系", 神経情報科学サマースクール NISS2001 講義録, 日本神経回路学会誌, Vol.9, No.1, pp. 36-40, 2002.

付録 A

実験タスクのトポロジー

図 A.1，A.2，A.3 に，本研究で実施した 3 種類の被験者実験におけるそれぞれのタスクのトポロジーを示す．

それぞれの実験タスクのトポロジーの特徴として，以下の点が挙げられる．まず，ある状態における行動は常に 2 種類あり，それぞれの行動によって遷移する状態が異なる．なお，自分自身への遷移や無限ループに陥ってしまうような遷移は存在せず，ゴール状態への遷移はある状態における特定の行動のみによる．また，初期状態からゴール状態までの最短経路が 2 種類存在し，その最短ステップ数は実験 1，2，3 のそれぞれのタスクにおいてそれぞれ 5，7，9 である．

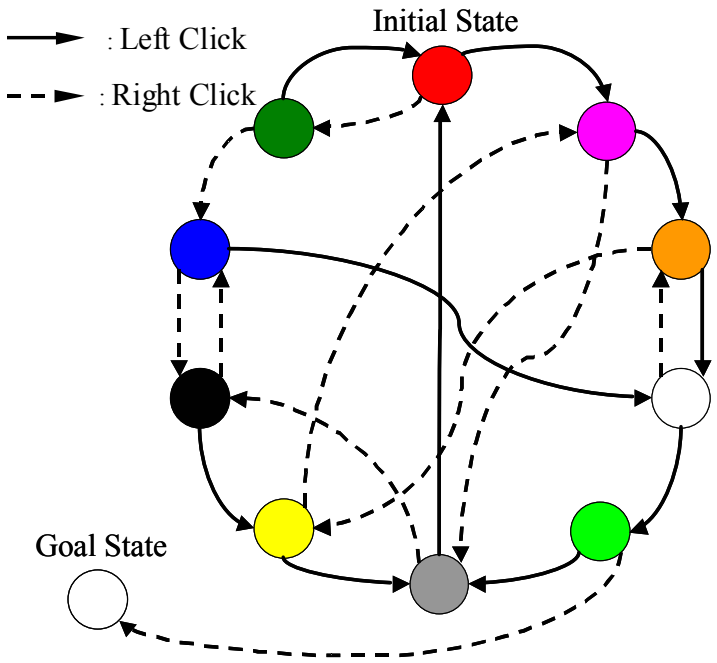


図 A.1　実験 1，タスク a のトポロジー

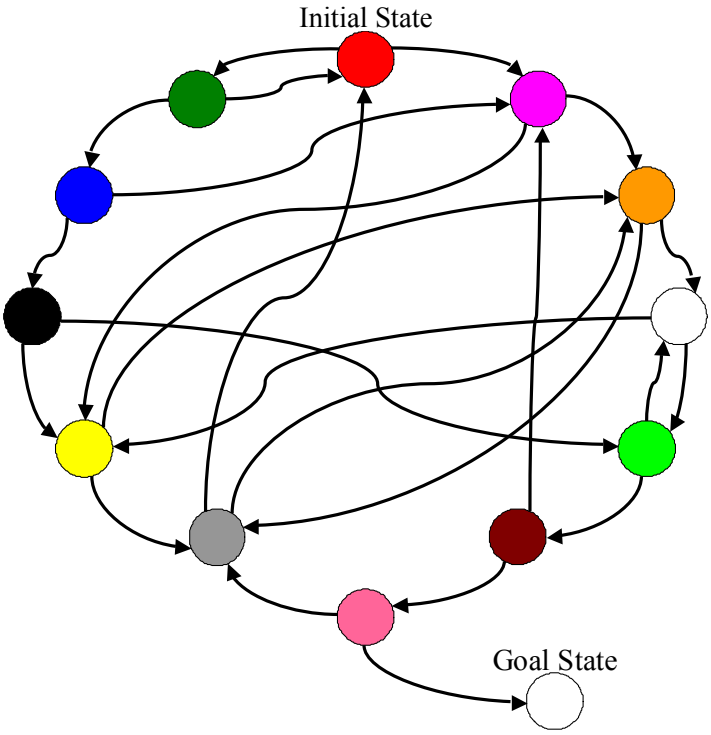


図 A.2　実験 2 のタスクのトポロジー

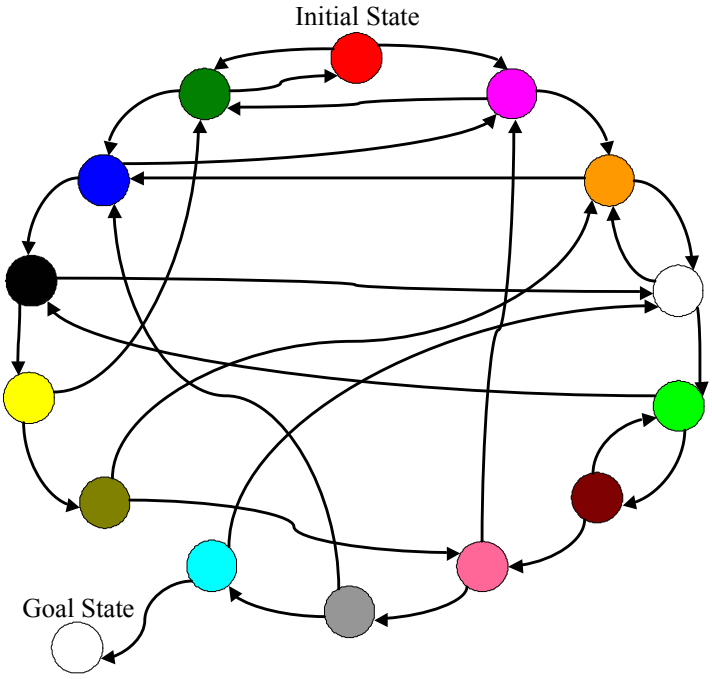


図 A.3　実験 3 のタスクのトポロジー