博士論文 令和 4 (2022) 年度

学習心理学と行動分析学における数理工学的手法の実践と意義

慶應義塾大学大学院社会学研究科

山田 航太

要旨

本博士論文では, 学習心理学と行動分析学という実験心理学における学習研究の分野において, 伝統的に扱われてきた現象に対し, 強化学習や深層学習といった計算論的手法を用いてアプローチすることを通して, 計算論的手法の導入による学習研究の可能性の拡張を目指した. 実験心理学における学習心理学と行動分析学の両分野においては, それぞれパブロフ型条件づけとオペラント条件づけという 2 つの条件づけを主題として, 様々な学習性の行動についての記述, 説明がなされてきた. 一方で, 機械工学の分野を中心に発展してきた強化学習は, これら両分野に影響を受けて誕生した機械学習の一領域である. 強化学習の枠組みには, 上記の 2 つの条件づけ, 及びそれらが関わる学習性の行動を包括的に扱う可能性が秘められている. 本研究では, パブロフ型条件づけとオペラント条件づけが関わるいくつかの行動現象を, 強化学習という枠組みの中で包括的に扱うことによって, 実験心理学における学習研究において新たな可能性を示した.

研究 1 においては, 瞬間的な行動と瞳孔の動態を, 計算論的な手法により計測および解析することを通して, 瞳孔の大きさが報酬に対する予測を反映すること, そして, 報酬の予測に伴う運動表出が薬物による阻害によって抑制されても, 瞳孔の拡大が生じることを明らかにした. 頭部固定装置を用いて, マウスにパブロフ型条件づけを行い, 音刺激によって報酬が予測できる群と, 報酬が予測できない群の 2 群をそれぞれ訓練し, その際のリッキングと瞳孔の大きさの変化を計測した. 報酬が予測できる群では, リッキングと瞳孔の大きさは, 音刺激によって上昇した. 次に報酬の予測に伴って生じるリッキングが瞳孔の大きさに対して与える影響を低減させるために, ドーパミン D2 拮抗薬であるハロペリドールをマウスの腹腔内に投与することで, リッキングを抑制した. 薬理学的操作の結果, ハロペリドールによってリッキングは濃度依存的に抑制されたが, 瞳孔の大きさの抑制は見られなかった. さらに, 瞳孔の大きさが報酬予測を反映していることを裏付けるために,

リッキングの動態を強化学習に基づいてモデル化することを通して, 報酬予測の動態を推定し, その情報をもとに瞳孔の大きさの時間的な変化が予測できることを示した.

研究 2 においては, オペラント反応の消去に伴って生じる, 消去バーストという現象の制御変数を, 強化学習モデルを用いたシミュレーションによって同定し, マウスの実験によってその妥当性を検証した. マウスのオペラント反応の瞬間的な動態とその背後にある内部過程を強化学習の枠組みに基づいてモデル化した. さらに, 提案モデルでは, 強化学習における好奇心というアイデアに着目し, 好奇心駆動型の強化学習をモデルとして実装することで, 従来の行動分析学と学習心理学では説明できない現象であった消去バーストが生じる環境条件を明確に予測できることを示した. さらに, 強化学習モデルによって明らかになった環境条件の下で, 実際のマウスの行動においても消去バーストが生じることを明らかにした.

研究 3 においては, オペラント反応の局所的なバーストと休止期間によって特徴づけられる, バウト・休止パターンを強化学習によってモデル化することを通して, 行動の背後のあるメカニズムを検証した. ここでは, 現実の動物が従事しうる実験とは関連のない行動, すなわち他行動を, 強化学習における状態として導入し, それらの間での遷移コストを仮定することにより, バウト・休止パターンという反応の時間構造を説明することを試みた. さらに, 他行動の存在と遷移コストのいずれかを欠損させたモデルではバウト・休止パターンが生じないこと, そして双方を実装したモデルでは, バウト・休止パターンのみならず, これまでの動物実験で蓄積された結果を再現できることを示した.

研究 4 においては, ネットワーク科学的な視点を強化学習に導入することで, 研究 3 において示した, 学習と行動の理解のための他行動という概念の導入というアイデアを洗練し, オペラント条件づけと, パブロフ型条件づけの双方が関わるとされる習慣形成と呼ばれる現象に対して, 新たな仮説を提唱した. ここでは, 行動を反応が相互結合したネットワークとみなすことで, 強化学習における状態遷移

に, ネットワークの構造という, 従来の行動分析学と学習心理学にない視点を導入した. この試みにより, 習慣形成と呼ばれる現象をネットワークの構造変化として説明できることを示した.

研究 1 ではパブロフ型条件づけによって獲得された反応の動態, 研究 2 ではオペラント反応の消去に伴う変化の過程, 研究 3 と 4 では, バウト・休止パターンと呼ばれるオペラント反応の時間構造や, オペラント反応とパブロフ型条件づけの双方が関与する習慣形成という現象を, 他行動に注目して, 強化学習によってモデル化した. このように強化学習は動物の学習性の行動を, オペラント条件づけとパブロフ型条件づけの双方を包括的に扱うことが可能であることを示した. 強化学習は神経科学や実験心理学内の諸領域でも応用されており, 今後はそうした隣接領域との接点を創出することも可能となる. 個々の研究では, 強化学習による機械論的, 数理的な行動, 内部過程の記述によって生み出される, 行動分析学と学習心理学の異分野との接続性についても考察した. 本博士論文においては, このような個々の実験事例を挙げながら, 計算論的手法が行動分析学と学習心理学という実験心理学における学習研究の既存の文脈の中で果たす役割と, 他領域との接点を作り出すという役割, つまりは行動分析学と学習心理学の深化と拡充のツールとしての意義を示した.

目次

要旨 2
目次 5
序論 9
学習をめぐる 3 つの学問 9
学習心理学とパブロフ型条件づけ 9
行動分析学とオペラント条件づけ 11
学習心理学と行動分析学の違い 13
機械学習が学習心理学と行動分析学にもたらすもの 14
強化学習について 17
強化学習とパブロフ型条件づけ 19
強化学習とオペラント条件づけ 22
計測技術と行動分析学・学習心理学の発展 23
本研究の位置づけ・目的 27
研究 1：瞳孔計測による報酬予測と報酬予測に基づく行動表出の分離 29
背景と目的 29
実験 1：報酬が予測可能な状況において瞳孔は拡大した 31
方法 31
結果 37
実験 1 考察 40
実験 2：運動表出を抑制しても瞳孔サイズは報酬予測的に拡大した 40
方法 40
結果 41

実験 2 考察 48
研究 1 総合考察 48
研究 2：好奇心駆動型強化学習による消去バーストの制御変数の同定 53
背景と目的 53
実験 1：シミュレーションによる消去バーストの制御要因の同定 56
方法 56
結果 59
実験 1 考察 64
実験 2：強化確率はマウスの消去バーストの有無を決定づけた 66
方法 67
結果 68
実験 2 考察 80
研究 2 総合考察 81
研究 3：強化学習によるバウト・休止パターンのシミュレーション 87
背景と目的 87
実験 1：選択とコストによるバウト・休止パターンの再現 91
モデル 91
シミュレーション 94
結果 95
実験 1 考察 98
実験 2：提案モデルによる過去の実験結果の再現 99
モデル 99
シミュレーション 99
解析 100

結果 100
実験 2 考察 106
研究 3 総合考察 107
研究 4：行動ネットワークの構造変化としての習慣形成 111
背景と目的 111
行動ネットワーク 113
実験 1：オペラント反応へのエッジの集中によって習慣形成が生じた 116
モデル 116
シミュレーション 118
結果 119
実験 1 考察 122
実験 2：提案モデルによる習慣形成に関わる要因の効果の再現 123
シミュレーション 124
結果 125
実験 2 考察 128
実験 3：モデルによる習慣形成を規定する要因の検討 129
シミュレーション 133
結果 133
実験 3 考察 136
研究 4 総合考察 137
研究 1−4 まとめ 143
総括 145
強化学習による分野間の相互乗り入れ 145
計測技術に期待されるもの 147

学習心理学と行動分析学における計算論的手法の意義..... 149
参考文献 150
関連業績 177

序論

学習をめぐる 3 つの学問

学習という用語を一意に定義するのは難しいが, 心理学においては「経験による行動の変容」という定義が広く用いられてきた (Hilgard and Bower, 1966; Mazur, 2016). 学習には大きく分けて 2 つの条件づけが含まれる. 1 つはパブロフ型条件づけ (あるいはレスポンデント条件づけ), もう 1 つはオペラント条件づけ (あるいは道具的条件づけ) と呼ばれる. この 2 つの条件づけは, 学習心理学と行動分析学という 2 つの心理学の学問領域で研究がなされてきた. そして, 学習を扱うもう 1 つの学問として, 機械学習, 特にその一領域である強化学習がある. 強化学習は, 学習心理学や行動分析学などの, 動物の学習を扱う心理学に影響を受けて誕生した, 機械学習の一種であり, 2 つの条件とそれらが関わる行動現象を包括的に扱える可能性を秘めている. まずは, 学習心理学と行動分析学, 及びそこで扱われてきたパブロフ型条件づけとオペラント条件づけについて紹介し, その後に強化学習を中心に, 機械学習が心理学の学習研究において, どのように貢献できるかを考察する.

学習心理学とパブロフ型条件づけ

パブロフ型条件づけは, ロシアの生理学者 Ivan P. Pavlov が条件反射として報告した現象に端を発する. Pavlov は, イヌにメトロノームの音を聞かせた後に, エサを与えることを何度も繰り返すことで, エサに対して観察されていた唾液分泌が, メトロノームの音に対しても観察されるようになることを報告した (Pavlov, 1927). メトロノームの音のように, 個体の反応を引き起こさない刺激は中立刺激 (neutral stimulus; NS), 生得的な反応を引き起こす, エサのような刺激は無条件刺激 (unconditioned stimulus; US), US によって引き起こされる反応は無条件反応 (unconditioned response; UR) と呼ばれ, NS と US を繰り返し対呈示することによって, NS に対しても条件反応 (conditioned response; CR) が生じるようになる.

このとき, CR を生じさせるようになった NS を CS (conditioned stimulus; CS) と呼ぶ. CS と US の対呈示によって, US に対して生じていた UR が, CS に対しても CR として生じるようになる現象, そして CS と US の対呈示という手続きをパブロフ型条件づけと呼ぶ. このパブロフ型条件づけに関する蓄積の多くは, 学習心理学によってなされてきた.

学習心理学は, 連合と呼ばれる概念を中心に据えて, 個体の行動を説明する実験心理学の一分野である. 連合とは, 刺激 (S), 反応 (R), そして結果 (O) の表象間のつながりを指す. 連合は, 刺激と反応 (S-R)から, 刺激と結果 (S-O), 反応と結果 (R-O) まで, あらゆる表象間に形成される (Rescorla, 1991). 例えば, オペラント反応にも, S-R 連合によって制御されるものと, R-O 連合によって制御されるものが存在する (Dickinson, 1985; Rescorla, 1991). Colwill and Rescorla (1985) は, ラットのレバー押しとチェーン引きのそれぞれを, 餌と水の異なる報酬によって訓練した. オペラント条件づけにより反応が獲得された後に, どちらか一方の報酬を味覚嫌悪条件づけによって低価値化させた. すると, ラットは低価値された報酬への反応を止めて, 低価値化されていない報酬と結びついた反応にのみ従事するようになった. これは反応が報酬の価値によって制御されていること, つまり R-O 連合によって制御されていることを意味する. その一方で Adams and Dickinson (1991) では, 1 つのオペラント反応を長期に渡って訓練することで, その反応が報酬低価値化の影響を受けなくなることを報告した. これは長期的な訓練によって, 反応の制御が R-O 連合から S-R 連合に移行したことを意味する (Adams and Dickinson, 1991; Dickinson, 1985). さらに S-O 連合の存在を示す研究として Pavlovian-instrumental transfer (PIT) と呼ばれる現象がある (Colwill and Rescorla, 1988). これはオペラント反応を, 任意の報酬によって訓練した後に, オペラント条件づけとは異なる事態で, 任意の CS とオペラント条件づけに用いた報酬でパブロフ型条件づけを行う. その後, オペラント条件づけ事態で, 報酬と結びついた CS を呈示すると, CS の非呈示時と比べて, より多くオペラント反応に

従事した. これは刺激と結果の間に連合 (S-O) が形成されたことを示唆する. これらの研究から刺激, 反応, そして報酬の, それぞれの表象間に連合が形成されると結論付けられる (Rescorla, 1991). このように, パブロフが示した最初期の例である, CS が US と対呈示されることで US に対する反応を喚起させる能力を獲得すること (Pavlov, 1927) は, 条件づけの中のごく一部である. 今はより多くの表象間のつながりとして, 拡張された形で理解されている (Rescorla, 1988). この連合によって, 動物の学習性の行動を扱う実験心理学の一分野が学習心理学である.

行動分析学とオペラント条件づけ

行動分析学は心理学の一分野であり, 所与の環境の下での, 個体の振る舞いを予測すること, あるいは個体の振る舞いを制御するために, いかに環境を整えるか, ということに重きを置いてきた. その研究の多くは, 環境上の変数と, 個体の行動の間にある関数関係を同定することに注力しており, 行動分析学は, 個体の行動と, それを制御する環境の変数を数多く明らかにしてきた.

行動分析学の最も核にあるアイデアは強化の原理であり, その発見は, Edward L. Thorndike の「効果の法則」の提案にまでさかのぼる. Thorndike は ”Animal intelligence”の中で, 自身の一連の実験を通して「効果の法則」という動物の学習法則を提案した (Thorndike, 1889). 彼の実験は問題箱という実験装置を用いて実施された. 問題箱の内部にはいくつかの仕掛けがあり, その仕掛けを解くこと, 例えば動物が装置内にある紐を引っ張るなどの反応, で装置から出られるように設計されていた. Thorndike は, この装置内に, 空腹にしたネコやイヌを入れ, 装置から外へ脱出するまでの潜時を計測した. そして, 何度も試行を繰り返すにつれて, 脱出までにかかる潜時が短くなっていくことを発見した. Thorndike はこの際の動物の行動の変化を詳細に記述している. 動物を装置に導入した最初の頃には, 箱の中で暴れるといった行動が観察された. しかし, そうした行動の一部が脱出につながることで, 徐々に箱からの脱出に寄与する行動が増加し, それ以外の行動が減少したことを報告している. このように, ある刺激の下で生じた反応の結果, その個

体にとって満足のいくものが得られた場合には, その刺激と反応とが結びつき, 同様の刺激の下で, その反応が生じやすくなる. これが「効果の法則」と呼ばれる学習法則である. 行動の増減は必ずしも動物にとって, 好ましい事象によってのみ引き起こされるわけではない. ある事象が行動に随伴して生じることで, 後の行動の頻度が増加する場合もあれば, 逆に減少する場合もある. 一方で, ある行動の後に, ある事象が消失することで, その行動が増加, あるいは減少することもある. このように, 行動に随伴する事象によって制御される行動をオペラント行動と呼ぶ. さらに行動に随伴し, その将来の生起頻度を増加させるようなものを強化子, 逆に生起頻度を減少させるようなものを罰 (あるいは弱化子) と呼ぶ. このように, Thorndike が「効果の法則」として提案した動物の学習法則は, 行動に随伴する事象の生起と消失, そしてそれによる行動の増減という 4 つの組み合わせへと一般化され, 強化の原理として定式化された.

行動分析学は, 強化の原理に基づいて, 個体の行動と, 環境の関数関係を同定してきた. Ferster and Skinner (1957) では強化スケジュールと呼ばれる, 様々な強化子の呈示規則の下での, 個体の行動を記述しており, 同一の強化スケジュールの下で, 個体間に共通した行動パターンが観察されることを発見した. 基本的な強化スケジュールには, 固定比率 (fixed ratio; FR), 固定時隔 (fixed interval; FI), 変動比率 (variable ratio; VR), そして変動時隔 (variable interval; VI) スケジュールがある. FR スケジュールでは, 任意の回数の反応が生じるごとに強化子が与えられる. 例えば, FR 10 という場合には, ラットのレバー押しが 10 回生起するごとに, 強化子が呈示される. FR スケジュールの下では, 強化子の呈示直後に, 強化後休止と呼ばれる無反応時間が生じて, その後, 次の強化子呈示までは一定のペースで反応が生じる. FI スケジュールでは, 一定の時間間隔で強化子が利用可能となる. 例えば, FI 10 秒という強化スケジュールでは, 直前の強化子呈示から 10 秒経過後の最初の反応に対して強化子が呈示される. この強化スケジュールの下では, 設定された時間間隔に向かって反応が徐々に増加する, スキャロップと呼ばれる反応パターン

が形成される. VI と VR スケジュールは, FR や FI のような一定の反応数, 時間間隔ではなく, ランダムな回数, 時間間隔で強化子が呈示される強化スケジュールである. どちらの強化スケジュールでも, 反応は一定のペースで生じるが, VR スケジュールでの反応率が, VI スケジュールでの反応率を上回ることが報告されている (Baum, 1993; Baum and Grace, 2020; Ferster and Skinner, 1957). 強化スケジュールには, こうした単純なものだけでなく, 複数の強化スケジュールを組み合わせた, より複雑なものも存在する. その 1 つに, 実験装置内に複数の反応対象を設置して, それぞれに異なる強化スケジュールを割り当てる, 並立スケジュールがある. 並立スケジュールは, 選択行動の研究で頻繁に採用され, 複数の選択肢に異なる強化率の強化スケジュールを割り当てた際に, 強化率の比と反応率の比が一致するマッチング法則 (Herrnstein, 1970) や, 選択から強化子呈示までの間に遅延時間を含む事態において, 選択後の遅延時間と選択パターンの関係などが明らかにされてきた (Ainslie, 1974; Fantino et al., 1993). このように行動分析学では, 様々な環境の下で, 環境と個体の行動の規則的な関係が明らかにされてきた.

学習心理学と行動分析学の違い

ここまでで, 学習心理学と行動分析学, そしてそこで扱われてきたオペラント条件づけとパブロフ型条件づけについて紹介したが, ひょっとすると, これら 2 つの分野の違いは, 扱う条件づけの種類だけに思えるかもしれない. もちろん, 行動分析学でもパブロフ型条件づけに関する研究もあれば, 学習心理学でもオペラント条件づけに関する研究はあり, 決して扱ってきた条件づけの種類が, これらの分野を隔てるものではない. それでは一体何が, この 2 つの分野を分かつのだろうか.

両者の違いは行動の説明の方式にある. 行動分析学では, 個体の行動の説明する上で, 客観的に観察, そして直接的に操作不可能な心的概念や媒介変数を導入することを拒み, 行動上の変動性と行動に随伴する事象による強化や弱化という, 結果による選択 (selection by consequence) によって, 行動の変容を説明しようと試みた (Skinner, 1981). 従って, 行動分析学では, 環境との相互作用の中で行動が

徐々に形作られていくとみなされる. 従って, 動物の内部に何らかの表象や, その表象間の連合といったものは想定しない. 一方で学習心理学では, 前章で説明したように, 動物の学習を, 連合という表象間のつながりによって説明するものであった. こうした行動の説明方式の違いは, 学習心理学と行動分析学を分かつ特徴の 1 つであり, その背後には, 学習心理学が方法論的行動主義, 行動分析学が徹底的行動主義, という異なる哲学的な立場が存在する (丹野, 2019).

こうした行動の説明方式の違いは一見すると相容れないようだが, 行動分析学でも, あらゆる心的概念や媒介変数を無暗に拒絶するわけではなく, 個体と環境との相互作用の履歴を縮約的に表現するような形での導入を肯定する向きもある (Staddon, 2014). 学習心理学の立場からも, 環境と行動の数理的に表すことによって, 行動分析学と学習心理学は接続可能であると主張されている (澤, 2021). このように, 数理的に環境と行動の関係を記述することで, 行動分析学と学習心理学は決して対立的な立場とはならない.それでは, 行動分析学と学習心理学で扱ってきた, オペラント条件づけやパブロフ型条件づけ, そしてそれらが関わる様々な行動現象を包括的に扱いうる枠組みはあるのだろうか.

機械学習が学習心理学と行動分析学にもたらすもの

ここからは, いくつかの具体的な例を紹介しながら, 強化学習をはじめとする計算論的手法が, 学習心理学と行動分析学において, どのような貢献をするか検討する. 1 つは, 強化学習による行動のモデル化を通した, 理論的な側面での貢献が挙げられる. そこには, 動物の行動を捉える上での多角的な視点を提供することや, モデル化という行動の抽象化による分野間の接続性の創出, モデルによって導き出される新たな現象や実験事態の予測, などが含まれる. もう 1 つは, 教師あり・なし学習による, 新たな指標や既存の指標の洗練といった計測面での貢献がある. 計測系の改良を通して, 従来では捉えることのできなかった, 動物の学習や行動の新たな側面を切り出すことができるようになる.

強化の原理や連合学習といった, 学習心理学・行動分析学の根幹にある原理は非常にシンプルながらも, 個体の行動を説明する上で一定の成功を収めており, それらは心理学外へも波及した. 強化学習は機械学習の一領域であり, 行動分析学をはじめとする, 動物の学習を扱う心理学からの影響を受けている. 強化学習では, エージェントが試行錯誤を通して, 報酬がより多く得られる行動を学習することが目的とされる (Sutton and Barto, 2018). この問題設定は, Thorndike の効果の法則から行動分析学にまで受け継がれた個体の学習観に基づいたものであり, それは表面的な類似性に留まらず, 行動分析学や学習心理学との高い親和性を持ち, 両分野で扱ってきた諸現象を包括的に扱える可能性を秘めている. さらに, 強化学習は神経科学や他の心理学領域でも応用されており, そうした諸領域との架け橋ともなりうる.

学習心理学と行動分析学との親和性が認められる一方で, 強化学習は神経科学とも強い接点がある. その契機となったのは, Schultz et al. (1997) の研究である. TD 学習 (Sutton and Barto, 2018) は, US, あるいは報酬への予測と実測値の差分, すなわち予測誤差を用いた学習アルゴリズムである. Schultz et al. (1997) では, パブロフ型条件づけ課題中の, サルの中脳のドーパミン細胞の活動が, TD 学習における報酬予測誤差と一致することが報告された. Schultz et al. (1997) が報告した神経活動を予測誤差による学習の教師信号として解釈することについては異論があるものの (Barter et al., 2015; Redgrave et al., 2010), 報酬予測誤差仮説は今に至るまで, 学習の神経基盤の中核として考えられている (Glimcher, 2011; Wise, 2004). さらに強化学習は, 学習だけでなく, 意思決定や時間知覚といった幅広い心理・神経科学のトピックへと応用されており, ヒトや動物の行動をモデル化するツールとして頻繁に採用されるようになりつつある (Dayan and Daw, 2008; Gershman and Daw, 2017; Niv, 2009; Petter et al., 2018).

こうした強化学習の多岐にわたる成功の一方で, 強化学習の誕生の背景にある行動分析学は, 環境と行動の関数関係に注目する上で, 個体内の心的過程や行動の

背後にある神経メカニズムを扱ってこなかった. これは, 行動の予測と制御, という観点において, 個体内部の過程が冗長であるとされるからである (Skinner, 1953, 1999). 例えば, 動物に水を飲ませるためには, のどの渇きを操作する必要がある. しかし, 現実的にのどの渇きを直接的に操作することはできず, 水の摂取する機会を操作することで, 個体に水を飲ませるように仕向けることとなる. 同様に, 個体の飲水量を予測する上で, のどの渇き具合は有用な予測子となるが, それを直接的に観察することはできない. このように, 個体の行動に介入するために, 環境上の変数を直接的に操作することは可能だが, 個体内部の変数を直接的に操作することや, 観察することはできない. しかし, 近年では強化学習やベイズ推定, 深層学習といった計算論的な手法と神経活動の計測を組み合わせることで, 行動や知覚体験のデコードや機械の制御, あるいは神経活動操作による行動の直接的な操作が可能となりつつある (Bernstein and Boyden, 2011; Carmena et al., 2003; Chapin et al., 1999; Kamitani and Tong, 2005; Levedev et al., 2005; Musk, 2019; Shams et al., 2000). こうした現状を鑑みると, 動物の内部で起きている事象は, 行動の予測と制御, という観点から見ても, 有用かつアプローチ可能なものとなりつつある. そして, 強化学習とはじめとする計算論的な手法は, 行動分析学と神経科学との接点となる可能性を秘めている.

計算論的な手法は, 神経科学との接点を生み出すだけでなく, 学習心理学や行動分析学で新たな検証可能な問いを生み出す可能性に溢れている. 前述の Schultz et al. (1997) の研究においては, TD 学習や Rescorla-Wagner モデル (Rescorla and Wagner, 1972) における予測誤差という計算上のパラメータが, 中脳のドーパミンと呼ばれる神経伝達物質を放出する神経細胞の活動によって担われていることが示され, 生物の脳内で学習理論に対応した計算処理が表現されている可能性が示された. それ以外にも R-W モデルの登場以降, モデルから予測された新たな現象の発見として, 過剰予期効果, 超条件づけといった現象が存在する (Miller et al., 1995). 強化学習でも, モデルのパラメータや内部パラメータと神経系との対応が

広く検討されるようになり, 神経活動の解釈や作業仮説の提供が行われてきた (Gläscher et al., 2010; Dabney et al., 2020; Doya, 2002; Samejima et al., 2005). このように計算論的な行動のモデリングは, 実験心理学で得られてきた知見から神経科学へと検証可能な仮説を提供することや, 実験心理学という学問領域の内部においても新たな現象の予測や仮説の提供において重要な役割を果たすことが大いに期待される.

さらに, 計算論的な手法は, 行動や神経活動のモデリングだけに留まらず, 計測系へも応用され, 従来の解析手法では困難であった高精度な動物の動作のトラッキングや, 実験中に生じる様々な反応の計測といった, 網羅的な行動計測を可能にしている. こうした計測系や指標の発展もまた, 実験心理・神経科学の裾野を広げる役割を担うようになっている. その意味で, 計算論的手法の導入は, 行動分析学が注力してきた, 行動の予測と制御のツールとして, 神経科学と実験心理学のインターフェースとして, そして学習と行動の研究に新たな視点を導入し, 検証可能な問いの裾野を広げるためのツールとして, 行動分析学や学習心理学が扱ってきた様々な学習および行動的現象の研究を前進させる可能性に満ち溢れている. 以降では強化学習や深層学習などの手法が利用された事例を紹介しながら, 行動分析学と学習心理学で, 計算論的手法を導入する意義について, より詳細に述べる.

強化学習について

強化学習の問題は, エージェントが所与の環境内で, 任意の報酬信号を最大化するために, 状態とそこで取るべき行動の対応を学習することである. 学習は, エージェント自身が試行錯誤を通して, 報酬がより多く得られる行動を発見することで行われる (Sutton and Barto, 2018). エージェントは, 学習者であり意思決定者である. 環境はエージェントが相互作用する, エージェント外部の全ての事象を含む. エージェントと環境の関係は, エージェントが行動によって環境に働きかけ, 環境がその行動に応じて, エージェントへ新たな状態と報酬を与える, というループによって表される. エージェントは環境と相互作用することを通して, ある状態

で得られる報酬, 状態遷移に関する予測, そして報酬をより多く獲得するための行動を学習する. ある任意の状態における報酬や状態遷移に関する予測を獲得することは, 動物の学習におけるパブロフ型条件づけの側面に, 予測に基づいて報酬をより獲得可能な行動を学習することは, オペラント条件づけ的側面として考えられる (鮫島, 2022). 以降では強化学習の基本的な枠組みと, いくつかの具体的なアルゴリズムを紹介することを通して, 強化学習におけるパブロフ型条件づけとオペラント条件づけの扱いについて述べる.

まずは, エージェントと環境の相互作用をより詳しく解説する. これらの相互作用は離散的なタイムステップとともに進行する. このタイムステップは特定の単位を持たないため, 実際の現象や実験手続きに合わせて設計する必要がある. タイムステップは実時間における任意の間隔と対応する必要はないため, 刺激の呈示や画面の遷移といった, 実験の段階として設計することも可能である. 各タイムステップにおいてエージェントは状態 $s_t \in S$ を観測する. S は所与の環境において可能な状態の集合を表す. 状態は具体的な実験場面での刺激の有無 (刺激なしをs_0 , 刺激ありを s_1)や, 複数の刺激があればそれぞれの刺激の呈示 (刺激数を N とすると状態空間は $S \in [1,2,3 \cdots N]$ となる)に対応する. エージェントは, その状態に基づいて行動 $a_t \in A(s_t)$ を選択する. $A(s_t)$ は, 任意の状態 s_t において, 可能な行動の集合を表す. 行動の集合 A は現実の生物が取りうる全ての行動を含む必要はなく, 実際の実験事態で可能な行動空間に制限される. 例えば, 選択場面であれば, 選択肢の数がそのまま取りうる行動となる. エージェントの行動選択後に, 次のタイムステップへ移行し, エージェントは環境から新たに状態 s_{t+1} を観測し, 報酬 $r_t \in R$ を受け取る. エージェントは与えられた報酬に基づいて方策 π の学習を行う. 方策は状態から行動への写像であり, 方策を学習することは, エージェントがより報酬が得られるような状態と行動の対応関係を学習することを意味する. 方策 π は行動の選択確率を表すものであり, 離散的な行動と連続的な行動の双方を含む. これらの要素から構成される強化学習では, 将来の状態や報酬の確率分布は, 現在

の状態と行動によってのみ決定され, それをマルコフ決定過程と呼ぶ. 従って環境のダイナミクスは $p(s', r \mid s, a) = Pr(s_{t+1} = s', r \mid s_t = s, a_t = a)$ と表される. 強化学習はこの抽象的な設計ゆえ, 様々な課題に適用することができる (Sutton and Barto, 2018). 古典的な課題ではバンディット課題や棒立てが使用され, 近年ではインベーダーやスーパー・マリオ・ブラザーズ, VizDoom, StarCraft, 碁というゲームにまで及び, これらは状態空間と行動空間が離散的なものから, 連続的なものまでをも含む (Ha and Schmidhuber, 2018; Mnih et al., 2015; Pathak et al., 2018; Silver et al., 2016; Vinyals et al., 2019). この強化学習の適用範囲の多様性は, 強化学習が学習心理学と行動分析学で用いられてきた実験手続き・現象を十分に射程に収めることができることを示唆する. 以降のセクションでは, 実際にパブロフ型条件づけとオペラント条件づけは強化学習の枠組み内で, どのように扱われているかを紹介する.

強化学習とパブロフ型条件づけ

強化学習におけるエージェントの目的は, 長期に渡る報酬を最大化するような方策を学習することであったが, それはどのように達成されるのだろうか. それは報酬や状態遷移に関する予測に基づいて, 報酬を最大化する行動を学習することである. そこで, ここからは, 強化学習における予測的側面について解説する.

まず, ある状態において, エージェントが獲得しうる収益を定義する必要がある. 収益をもっとも単純に表す方法は, 現在から将来に渡る報酬の合計値 $G_t = R_{t+1} + R_{t+1} + R_{t+2} + \cdots + R_T$, とするものである. この定義には明確な欠点が存在する. ここで T は最後のタイムステップを示すが, 最終ステップを定義することが困難な場合, 上記の形で収益を定義した場合に G_t が発散してしまう. そこで, 現在から時間的に遠い報酬ほど価値を割り引く, という割引率の概念が導入される. 割引報酬和は $G_t = R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \cdots = \sum_{t=1}^{\infty} \gamma^t R_{t+1}$ と表される. ここで割引率 γ は $0 \leq \gamma \leq 1$ の範囲を取るパラメータであり, $\gamma < 1$ とすることで, 現在から非常に遠い未来の報酬は, 実質的に無視されることとなる. さらに強化学習の問題

には, 方策と状態遷移という確率的な要素が含まれる. そこで割引報酬和を状態遷移確率と方策によって期待値を取ることで状態価値関数を定義する. 状態価値関数は $V(s) = \mathbb{E}_\pi[G_t \mid s]$ であり, ここで G_t は割引報酬和を表す. 従って $V(s) = \mathbb{E}_\pi[\sum_{k=0}^{\infty}\gamma^k R_{t+k+1}|s_t = s] = \mathbb{E}_\pi[R_{t+1} + \gamma\sum_{k=0}^{\infty}\gamma^k R_{t+k+2}|s_t = s]$ となる. ここで R_{t+1} と $\gamma\sum_{k=0}^{\infty}\gamma^k R_{t+k+2}$ を明示的に分けることで, 前者を即時報酬として, 後者を次の時点における状態 s' における割引報酬和とすることができる. それを状態遷移と方策によって期待値を求めると, $V(s) = \sum_a \pi(a|s)\sum_{s',a} p(s', r|s, a)[r + \gamma\sum_{k=0}^{\infty}\gamma^k R_{t+k+2}]$ となり, ここで割引率が掛かる項は次の時点における状態価値関数であることから, 状態価値関数は最終的に $V(s) = \sum_a \pi(a|s)\sum_{s',a} p(s', r|s, a)[r + \gamma V(s')]$ と表すことができる. 強化学習では, この状態価値関数によって, 現在の状態における報酬の予測が定義される.

それでは, 具体的に状態価値関数をどのように学習するのだろうか. 状態価値関数を学習する代表的な手法の 1 つに temporal difference (TD) 学習がある. この手法では, エージェントが自身の経験に基づいて, 状態価値関数を学習する. 具体的には, 各タイムステップにおける状態と報酬, そして次の時点の状態における状態価値関数の 3 つを用いて, 逐次的に状態価値関数を更新する. まず, 各タイムステップにおいてエージェントは TD 誤差, $\delta = r_t + \gamma V(s') - V(s)$, を計算する. ここで計算された TD 誤差を用いて, その状態における状態価値関数を, $V(s) \leftarrow V(s) + \alpha\delta$, によって更新する. ここで α は学習率であり, 行動価値関数の更新の幅を決定するパラメータである. TD 学習は以上の式に従って, 各タイムステップにおいて TD 誤差を計算し, 任意の学習率 α でもって, 徐々に状態価値関数を学習する手法である.

ここまでで, 状態価値関数の定義をして, その具体的な学習方法を紹介したが, ここで一度, 話を学習心理学に戻そう. TD 学習は連合強度の更新則として提案された R-W モデル (Rescorla and Wagner, 1972) と類似している. R-W モデルは CS と US の対呈示によって CS-US 間に連合が形成されると仮定し, その連合強

度の更新則を $V_{t+1}^{i} = V_t^i + \alpha\beta \cdot \left(\lambda - \sum_j V_t^j\right)$ と表した. ここで, α は CS の明瞭度, β は US の明瞭度, λ は獲得される連合の最大値を表す. TD 学習と比較してみると, TD 学習における学習率 α は R-W モデルにおける $\alpha\beta$ に対応する. 次に報酬 r_t は λ に対応する. 最後に, 状態価値関数 $V(s)$ は連合強度の総和 $\sum_j V^j$ に対応する. R-W モデルは試行ごと, つまり一回の CS と US の対呈示ごとの連合強度の更新則を表したモデルであり, TD 学習は状態遷移を考慮することで, 離散的な試行ではなく, 時間構造をモデル化しており, R-W モデルの時間的拡張ともみなせる (鮫島, 2022).

TD 学習は動物のパブロフ型条件づけ事態へと適用されており, 既に紹介した Schultz et al. (1997) がその代表例である. Schultz et al. (1997) では, パブロフ型条件づけ課題中の, サルの中脳のドーパミン細胞の活動が, TD 学習によって計算される報酬予測誤差と一致することが報告された. CS によって報酬ができない場合には, ドーパミン細胞は報酬に対して一過性の応答を示したが, 学習が進み, CS によって報酬が予測できるようになるにつれて, その一過性の応答は, 報酬に対してではなく CS に対して生じるようになった. さらに, CS によって報酬が予測できるようになった後に, CS のみを呈示して報酬の呈示を省略すると, ドーパミン細胞は, 本来の報酬呈示のタイミングに抑制された. このドーパミン細胞の活動は TD 学習における, 予測誤差に対応する. このように, TD 学習はパブロフ型条件づけのモデルとして, そしてパブロフ型条件づけの神経基盤を明らかにする立役者としての役割を果たした.

強化学習におけるパブロフ型条件づけの側面について, 基本的な定義から始め, 具体的なアルゴリズム, 学習心理学におけるモデルとの関連, そして実際の応用事例を紹介してきた. しかし, パブロフ型条件づけには, 様々な連合構造が存在することを思い出してほしい. ここまでで, 紹介したのは, 状態から報酬の予測に関する問題だけであり, それはパブロフ型条件づけのごく一部でしかない. 既に紹介した TD 学習は, 実際のエージェントが観測した状態遷移や報酬に基づいて状態価値

関数を逐次的に更新する手法であった. こうした手法のほかに, 状態遷移や報酬関数を直接的にモデル化し, それを学習させることで, 状態価値関数を学習する手法がある. TD 学習のような, 環境のモデルを陽に扱わない手法をモデルフリー学習, そして環境を陽にモデル化する手法をモデルベース学習と呼ぶ. モデルベース学習では, 状態遷移や報酬関数を直接学習するため, ある刺激の後に, どの刺激が出現するか, あるいは, ある行動によってどのような刺激, 報酬が得られるか, といったことを学習する. これは, 学習心理学における, S-S 連合や R-O 連合として解釈できる. ここでは手法の詳細については省略するが, そのいくつかの例を簡単に紹介する. 動物は, 自身がおかれた空間を探索することで, 認知地図を形成すると考えられている (Tolman and Honzik, 1930). Russek et al. (2017) では, 迷路をグリッド状の離散的な状態空間とみなし, Dyna-Q と呼ばれる, モデルベース学習によって, 状態間の遷移構造を直接モデル化, 学習することで, 認知地図の形成を示唆する潜在学習を再現している. 習慣形成における報酬への感受性の変化を, モデルベースな制御からモデルフリーな制御への移行として説明する (Daw et al., 2005), といった応用事例がある. このように, 強化学習では, パブロフ型条件づけの最も基本的な CS と US の対呈示に伴う予測の更新をはじめ, S-S 連合や R-O 連合といった, 様々な連合構造を取り扱うことができる.

強化学習とオペラント条件づけ

前章では予測という部分に着目して, 強化学習とパブロフ型条件づけの関連を紹介してきたが, ここからはオペラント条件づけに相当する, 報酬を最大化する行動の学習という問題を扱う. 状態価値関数が, 現在の状態における将来の報酬の予測を示すように, ある状態において, 任意の行動を選択することによって得られる将来の報酬も同様に定義することができる. それは行動価値関数と呼ばれ, $Q(s,a) = \sum_{s'} p(s', r \mid s, a)[r + \gamma V(s')]$ と定義される. この行動価値関数は状態価値関数と同様に, エージェントの経験から学習することができ, TD 誤差を $\delta = r_t + \gamma Q(s', a') - Q(s, a)$ として, 任意の学習率で更新することで学習することができる.

この行動価値関数は, ある状態において可能な行動の数だけあり, 実際の行動選択確率は softmax 関数やϵ-greedy 法といった, 任意の関数によって算出される. このようにして, 行動価値関数を学習することで, 報酬を最大化する行動を学習することができる.

オペラント条件づけ事態で報告されている行動現象に対しても, 強化学習は応用されている. 複数の選択肢が存在する自由オペラント事態において, 動物の反応率の比が強化率の比と一致するマッチング法則というものがあった (Herrnstein, 1970). Sakai and Fukai (2008) では, Actor-Critic learning という手法によって, 並立 VI VI スケジュールと並立 VR VR スケジュール下でのマッチング法則が再現できることを示した.さらに, こうした行動に関連した価値表象が線条体で報告されており, 価値学習やそれに基づく行動選択, そしてその背後にある神経メカニズムの解明へと繋がっている (Samejima et al., 2005).

ここまでで, パブロフ型条件づけとオペラント条件づけの研究で, 報告された現象への強化学習の応用事例と, その神経基盤に関する具体的な研究を紹介することで, 強化学習における 2 つの条件づけの扱いを説明してきた. これらの事例は, 強化学習が 2 つの種類の条件づけの双方を統一的に扱える可能性を秘めていることを示唆する. 本研究では, 行動分析学と学習心理学で扱われてきた, いくつかの行動現象へと強化学習を応用する. さらに, 行動をモデル化する時間スケールに着目することで, 強化学習という枠組みが, 2 つの条件づけとそれらが関わる行動現象を様々な時間スケールで包括的に捉えられることを示す.

計測技術と行動分析学・学習心理学の発展

強化学習は, 学習心理学や行動分析学に対して, 理論的な側面から貢献することが期待されるが, 機械学習の応用は計測系にも期待できる. 学習心理学と行動分析学の歴史を振り返ってみると, これらの学問の進展は実験装置や計測技術の発展とともにあった. 前章でも紹介した Thorndike の問題箱はその最初期の例である. Thorndike は ”Animal intelligence” 発表当時, Romanes らが用いていた様々な動

物の「知的な」エピソードを集めて, 解釈するアプローチに対して批判的であった. そこで, 逸話収集に代わって実験的な手法を用いること, そして連合という低次なプロセスで, 動物が見せる, ある種の知的な行動, 例えば推論や模倣などを説明しようと試みた (Thorndike, 1898). 実験室内で行われる実験だけが動物の行動を明らかにする唯一の方法ではないが, Thorndike が問題箱という装置を開発したことによって, 実験的な動物の心理学が始まったことは, 学習心理学と行動分析学の発展における装置や計測技術の開発の重要性を語る上で印象的な出来事であった.

学習心理学と行動分析学で, 最も使用されている実験装置は Skinner によって開発されたオペラント箱だろう. オペラント箱の開発の最初の試みは, 実験の自動化のためであった. Thorndike の問題箱や, 学習研究で用いられていた迷路のような実験装置は, 試行ごとに動物をスタート位置へ移動させることや, 箱の中に再導入するなどの手間が生じる. Skinner はまずラットの走路を用いた実験で, ラットが自身で開始位置に戻れるような仕掛けを作成し, エサを自動で供給し, 反応を経時的に記録する累積記録器を開発する. こうした走路での, 実験の自動化の試みは, 最終的に箱の中に挿入されたレバーやキーへと置き換えられ, 現在のオペラント箱となる (Skinner, 1956). こうした開発された実験装置と, 累積記録という新たな指標は, 行動分析学の発展に寄与するとともに, 学習心理学や神経科学をはじめ幅広い分野へと普及することとなった.

こうした実験装置の発展だけでなく, 学習心理学と行動分析学の研究では計測装置にも工夫がされてきた. 例えば Guthrie は学習において反応と反応の間の連合を重要視しており (Guthrie, 1930), ネコの問題箱の実験で, ネコが反応する瞬間の写真をカメラで撮影することで, 徐々に反応の形態が固定化されることを報告した (Guthrie, 1946). 同様に, Jenkins and Moore (1973) では, ハトの自動反応形成中に, ハトがキーを突くたびにその瞬間の写真を撮影し, キー突きの反応形態を分析することで, 報酬の種類によって, 反応形態が異なることを報告した. これらの

手法は, 画像を用いて動物の反応の形態を分析する現代的な手法の先駆けであり, 学習に伴って反応が固定化するというアイデアや, 報酬による反応形態が異なるという仮説を検証可能な問いへと落とし込むことに成功した. こうした試み以外にも, オペラント箱の床下にマイクロスイッチを仕込むことによって, 活動量の定量化を行い, 強化子の呈示によって, 反応の強化にだけでなく, 一般的な活動量の上昇が明らかになった例 (Killeen, 1975) や, オペラント箱内での, 位置ごとの滞在時間の計測し, それに基づいた実験制御によって, マッチング法則が離散的な反応だけでなく, 反応の時間配分にも適用できることを示した例がある (Baum and Rachlin, 1969).

新たな指標の導入が学習の理論研究へと貢献した例として, 学習における注意の役割を明示した Kaye and Pearce (1987) の研究も挙げられる. この研究の紹介に入る前に, それ以前の学習の過程の捉え方を簡単にまとめる. Pearce-Hall モデル (Pearce and Hall, 1980) や Mackintosh の注意理論 (Mackintosh, 1975) で, 学習における注意の役割が明示的にモデル・理論化される以前のパブロフ型条件づけの代表的なモデルは R-W モデル (Rescorla and Wagner, 1972) であった. このモデルでは, 個体の学習速度を決めるパラメータは 2 つあり, それは CS の明瞭度 α と US の明瞭度 β であった. これらのパラメータは, 個体や実験操作によってその推定値が異なるものの, 時間的, あるいは学習に伴う変化は仮定されておらず, 一定とされていた. このような仮定では説明できない現象として, 潜在制止がある (Lubow and Moore, 1959). 潜在制止は, CS と US を対呈示する前に, CS だけを単独で何度も呈示する手続きであり, その結果として CS-US の対呈示で CR の獲得が遅れるというものである. R-W モデルでは α が一定であるため, CS 単独呈示による学習の遅延を説明できない. そこで Kaye and Pearce (1987) は潜在制止において, 個体が CS に対して向ける注意が減少していると仮定して, 注意を定義するために CS 呈示中の定位反応を計測した. そして, CS の単独呈示を繰り返すごとに CS への定位反応が減少することを報告した. これらの一連の研究は, 定位反応

という新たな行動指標の計測によって, 学習心理学に注意という心的概念を導入することに成功した例である. このように計測系や実験装置の発展は行動分析学と学習心理学の発展を支えてきた.

近年では, 深層学習のような機械学習の発展によって, より高精度かつ網羅的に動物の行動が計測可能となりつつある. こうした背景に生まれた, 計算論的エソロジーは, 計測にかかる時間や労力といったコスト, 観察者の主観性の問題, そして人の観察では見逃す行動の発見といった目的のため, 機械学習などのツールを, 動物の行動の解析と計測に応用する領域である (Anderson and Perona, 2014). 具体的には, 動物を撮影した動画を元に, 画像解析で動物の骨格推定などを行い, その骨格データを主成分分析などによって低次元ベクトルへと変換し, 時間構造をモデル, あるいはデータに埋め込むことで, 姿勢のダイナミクスを分類する. それにより, 動物の行動を網羅的に計測することが可能となる. この網羅的な行動計測によって, 薬理処置による反応間の遷移の影響を分析することや (Wiltschko et al., 2020), 特定の行動シーケンスに対応する神経活動の発見 (Markowitz et al., 2018) へと繋がっている.

こうした近年の計測技術の発展は, 学習心理学や行動分析学への応用が大いに期待される. 例えば, 行動分析学では, 他行動というオペラント反応以外の反応を含む行動クラスが仮定されることがある. 例えば, 単一スケジュール下での反応率と強化率をマッチング法則によって説明した Herrnstein (1970) では, 単一スケジュール下での動物の行動を, オペラント反応とそれ以外の反応との間での選択行動として捉えた. さらにバウト・休止パターンと呼ばれる, オペラント反応の局所的な反応のバーストとそれに続く休止期間によって定義される時間的構造も, こうしたオペラント反応と他行動との間の選択によって説明される (Yamada and Kanemura, 2020). 一方で, 他行動が計測されることは稀であり (稀な事例として Staddon and Simmelhag, 1971), ともすれば根拠に乏しい仮説構成体ともなりかねない. しかし, 計算論的エソロジーで用いられている手法を応用することで, 他

行動を網羅的に計測することが可能となる. こうした他行動の計測は, 上記のバウト・休止パターン以外にも迷信行動, スケジュール誘導性行動, そして反応形成など様々な現象に適用できることが指摘されている (松井, 2021). 既にある試みでは, 固定時隔スケジュールと変動時隔スケジュールの間での行動の違いを, 計算論的エソロジーで用いられるような手法で分析した研究も存在している (Leon et al., 2021). このように, 計測面での計算論的手法の活用によって, 動物の新たな学習や行動の側面を捉えることを可能とすることで, 学習心理学と行動分析学の発展が後押しされることが期待される.

本研究の位置づけ・目的

本研究では, 計測と理論の双方から, 計算論的手法を取り入れることで, 学習心理学で扱われてきた現象を, 包括的に捉えることに取り組む. それを通して, 従来の手法での限界の克服や, 新たな視点の導入による既存の研究の再解釈, 新たな問い立ての創出, そして神経科学との接点の構築を試みる.

研究 1 では, 瞬間的な行動と瞳孔サイズの動態を, 計算論的な手法により計測, 解析することで, 瞳孔サイズが報酬に対する予測を反映すること明らかにする. そして, それは薬理処置によって運動表出を抑制してもなお生じることを明らかにする.

研究 2 では, オペラント反応の瞬間的な動態を強化学習によってモデル化することを通して, 消去バーストの制御要因を明らかにする. 従来の消去を扱う理論では, 消去を行動や行動を制御する潜在過程の単調減少として捉えていたため, 消去によって生じる一過性の反応の上昇, すなわち消去バーストを説明できない. そこで, 消去を単調減少する単一の過程ではなく, 強化学習における好奇心というアイデアに着目することで, 消去の開始直後に一時的に上昇するもう 1 つの潜在過程を導入し, それら 2 つの過程の和として, 消去という現象を捉え直した. そして, モデルを用いたシミュレーションによって, 消去バーストの制御変数の同定を行う.

そこで得られた結果を基に, マウスによる実験を行い, モデルと, その予測の妥当性を検証する.

研究 3 ではバウト・休止パターンという, 反応の一過性のバーストとそれに続く休止期間によって特徴づけられる, オペラント反応の時間構造を強化学習によってモデル化を行い, シミュレーションによってモデルの妥当性を検証する. ここでは動物の行動をオペラント反応と他行動という, 2 つの状態間の遷移と捉えることで, バウト・休止パターンという数秒から数条秒という時間スケールで観察される時間構造を説明できることを示す.

研究 4 では, 研究 3 で提案した, 動物の行動をオペラント反応と他行動間の遷移とする見方を拡張する. 動物の行動を他行動とオペラント反応間の状態遷移から, 無数の反応が相互結合したネットワークとみなすことで, 行動のマクロな構造をモデルに組み込む. それにより, 習慣形成と呼ばれる現象を, 行動ネットワークの構造変化として説明できることを, シミュレーションによって示す. さらに新たに提案したモデルによって, 従来の理論とは相互排他的な実験事態を提案する.

本博士論文では, これらの研究を通して,行動分析学と学習心理学で扱ってきた現象を, 包括的に扱うことができることを示すと同時に, 計測や理論的な側面での計算論的手法を取り入れることで, 動物の学習と行動の新たな側面を明らかにする. そして計算論的手法が, 学習心理学と行動分析学にもたらす貢献について考察する.

研究 1：瞳孔計測による報酬予測と報酬予測に基づく行動表出の分離

背景と目的

研究 1 では, 深層学習を用いた計測系のアップデートによって, パブロフ型条件づけ中のマウスの瞳孔計測を, そして強化学習による行動のモデリングによって, 課題中のマウスの内的な報酬予測の推定を行うことで, 瞳孔サイズが報酬予測を反映していることを明らかにした. 従来では技術的な障壁のため, げっ歯類の実験中の瞳孔計測は容易ではなかったが, 頭部固定装置と画像解析によって, その問題を克服することに成功した. マウスの瞬間的な反応の動態を, 強化学習の枠組みでモデル化することによって, 報酬予測の動態を推定することが可能とした. そしてこの 2 つを結びつけることで, パブロフ型条件づけ課題中のマウスの瞳孔が, 報酬予測を反映していることを明らかにした.

将来の出来事を予測することは, 個体が報酬を獲得すること, 嫌悪的な出来事を回避することに役立つ能力である. パブロフ型づけは, そうした動物の予測的な能力を研究するために採用される実験手続きである. 初期のパブロフによる条件反射研究から, 唾液分泌をはじめ, 皮膚電気反応, 心拍, そして瞳孔といった生理指標もパブロフ型条件づけによって, 準備的な反応として獲得されることが報告されている (Esteves et al., 1994; Leuchs et al., 2017; Lonsdorf et al., 2017; Notterman et al., 1952; Öhman et al., 1976; Ojala and Bach, 2020; Pavlov, 1927; Pietrock et al., 2019; Wood and Obrist, 1964).

パブロフ型条件づけにおける瞳孔径の使用は半世紀以上前にさかのぼるが, 近年, その指標としての信頼性が見直されている (Finke et al., 2021). ヒトの恐怖条件づけや食餌性の条件づけにおいて, 条件刺激に対する条件反応として, 瞳孔の拡大が報告されている (Leuchse et al., 2017; Lonsdorf et al., 2017; Pietrock et al., 2019; Ojala and Bach, 2020). 瞳孔の大きさと, TD 学習における予測誤差 (Sutton

and Barto, 2018)や, Pearce-Hall モデル (Pearce and Hall, 1980) における刺激への注意といった学習理論との関係も議論されている (Koenig et al., 2017; Pietrock et al., 2019; Vincent et al., 2019). 瞳孔の大きさの変化は, 学習の文脈以外でも, 覚醒度, 注意, ワーキングメモリ, 社会的警戒, 選択課題における選択肢の価値, 不確実性など, 様々な内的状態と関連している (Ebitz et al., 2014; Ebitz and Platt, 2015; Finke et al., 2021; Joshi and Gold, 2020; Larsen and Waters, 2018; Van Slooten et al., 2018; Vincent et al., 2019; Zénon, 2019). これらの知見は, 瞳孔の大きさが条件刺激に対する条件反応であるだけではなく, 予測に影響を与える感覚運動処理の能動的な調節因子であることを示唆している (Ebitz and Moore, 2019).

行動の神経生物学的な基盤を理解する上での有用性にもかかわらず, げっ歯類の研究では, わずかな研究でしか瞳孔サイズを計測する試みがなされていない (Cazettes et al., 2021; Lee and Margolis, 2016; Nelson and Mooney, 2016; Privitera et al., 2020; Reimer et al., 2014; Wang et al., 2022). その原因には 2 つの技術的な問題がある. 第一に, 従来のげっ歯類の実験装置では, 被験体は自由に実験装置内で動くことができるため, 瞳孔を記録することが不可能であった. 第二に, こうした身体の運動は瞳孔サイズに影響を与えることが報告されている (Cazettes et al., 2021; Nelson and Mooney, 2016). それにより, ヒトの実験参加者と比較して, げっ歯類で瞳孔サイズを計測することが困難であった. 近年の実験装置と機械学習の発展によって, これらの技術的な限界を克服することが可能となった. 例えば, 頭部固定装置と DeepLabCut (Mathis et al., 2018; Nath et al., 2019) などの画像解析技術を組み合わせることで, 行動課題中のマウスの瞳孔や目の開きの定量化が可能となっている (Kaneko et al., 2022; Privitera et al., 2020).

本研究では, 頭部固定装置を用いて, パブロフ型条件づけ課題中のマウスのリッキングと瞳孔の反応の動態を調べた. 実験 1 では, 音刺激の呈示直後にスクロース溶液を報酬として与える延滞条件づけを, 頭部固定マウスに訓練して, リッキングと瞳孔の反応を記録した. この課題では, 報酬の予測可能性を操作するために 2 つ

の群を設けた. 随伴群では音刺激の直後に必ず報酬が呈示されることで, 音刺激によって報酬の到来を予測することができた. 非随伴群では, 音刺激と報酬の呈示がランダムかつ独立に行われるため, 音刺激は報酬の到来に関する情報をもたらさない. この報酬の到来が予測可能な群と予測不可能な群の間で, リッキングと瞳孔の反応の動態を比較した. 実験 2 ではドーパミン D2 受容体拮抗薬のハロペリドールを, 実験前に腹腔内投与することで, リッキングを抑制し, その時の瞳孔の動態を調べた. ハロペリドールは, リッキングや自発運動を抑制することが報告されており, 身体運動が瞳孔サイズに与える影響を低減することが可能となる (Arruda et al., 2008; Bernardi et al., 1981; Conceição and Frussa-Filho, 1996; Fowler and Mortell, 1992; Liao and Ko, 1995; Strömbom, 1977). さらに瞳孔サイズがマウスの報酬予測によって拡大することを裏付けるため, マウスのリッキングの瞬間的な動態を強化学習の枠組みモデル化して, そこから推定される報酬予測の動態から, 瞳孔サイズを予測できるかを検討した.

実験 1：報酬が予測可能な状況において瞳孔は拡大した

方法

被験体

被験体として 16 個体のオスの成体マウスを使用した. 全ての個体は実験開始時に過去の実験履歴はなかった. マウスは, 実験外ではホームケージで 12 時間の明暗期の下で飼育された. 全ての実験は暗期間中に行った. 実験期間中には給水制限を行い, 実験内でのみ水分を摂取するようにしたが, 日毎に体重を測定し, 必要に応じて追加での給水を行った. エサはホームケージ内で自由に摂食できた.

手術

手術では, マウスを 1.0～2.5%の割合で室内の空気と混合させたイソフルランにより麻酔を行い, stereotactic frame (942WOAE, David Kopf Instruments, Tujunga, CA) に固定した. その後, 実験中にマウスを頭部固定できるように, 頭蓋

にヘッドプレートを埋め込んだ. 手術後から実験開始までに 2 週間以上の回復期間を設けた.

手続き

手術後の回復期間を終えた後に, 48 時間の給水制限を行った. 実験初日の前日に実験装置への馴化を行った. 馴化では, マウスを実験装置内で頭部固定を行い, 口の前に金属製の吸口を設置して 10%のスクロース溶液を呈示した. 目視による確認によって, マウスが吸口からスクロース溶液を安定して舐めることができるようになるまで訓練を継続した. その後, パブロフ型条件づけで CS として使用する 6000 Hz の純音を 80 dB でランダムに呈示して CS への馴化を行った. マウスは, トンネル状に覆われたプラットフォーム上で, 頭部に外科的に埋め込まれたヘッドプレートを 2 つのクランプで挟むことで固定された. クランプはプラットフォーム横のスライドバーに設置されており, 個体ごとに適切な高さに調整した. プラットフォームの床は銅製のメッシュシートが敷かれていた. メッシュシートと吸口にはタッチセンサーが接続されており, マウスが吸口を舐めることにより, それらが通電することによって, リッキングを検出した. CS はプラットフォームの前方 30 cm に設置したスピーカーより呈示された (図 1.1.1. A).

馴化の後に, パブロフ型条件づけの訓練を行った. 実験条件は随伴群と非随伴群の 2 群を設け, 各群に 8 個体ずつマウスを割り当てた. 随伴群では CS がランダムな時間経過した後に, 1 秒間呈示され, その直後に 4 µl の 10% スクロース溶液が呈示された (図 1.1.1. B). CS の刺激間間隔は平均を 15 秒とする 10 – 20 秒の範囲からランダムに決定された. 非随伴群では CS と報酬がランダムかつ独立に呈示された (図 1.1.1. C). CS, 報酬の刺激間間隔は, 平均を 15 秒とする 10 – 20 秒の範囲からランダムに決定された. 1 セッションは両群とも 120 試行として, 訓練は 8 日間行った. 刺激呈示, 反応と動画の記録は python3 (3.7.8) で記述されたプログラムによって行われた. 実験は外部からの音を遮断するために, 実験室内で 75 dB のホワイトノイズを呈示し, 遮音箱内で行った.

瞳孔径計測

マウスの瞳孔を記録するため, 赤外線カメラによって課題中のマウスの頭部の動画を撮影した. カメラをマウスの正中線から 45°, 頭頂から 45 mm の位置に設置した (図 1.1.1. A). 実験室の明るさは 15 lux とした. 上記の条件の下で 30 FPS で動画の撮影を行い, 動画から瞳孔サイズの計測を行った. 深層学習によるトラッキングソフトウェア DeepLabCut によって, 瞳孔の縁を 8 点トラッキングし, その 8 点に対して楕円を当てはめた. 当てはめた楕円の面積を計算することで, 瞳孔サイズを計測した (図 1.1.1. D).

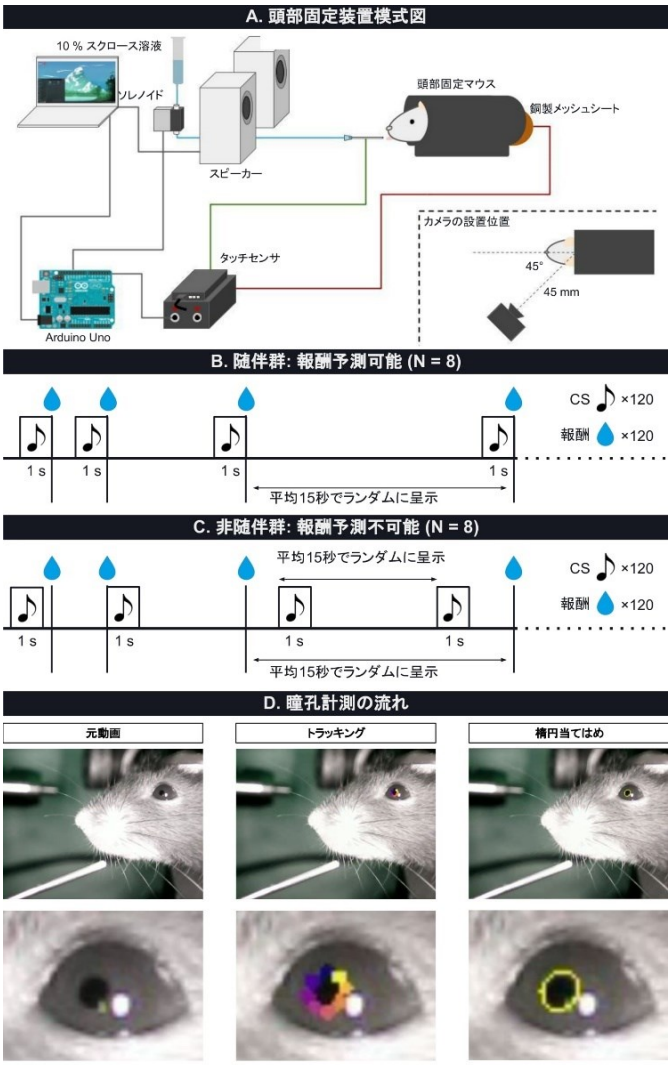

図 1.1.1. 実験装置・手続き・瞳孔計測方法の模式図.

A. 頭部固定装置と実験制御系の模式図. B. 随伴群の実験手続きの模式図. 1 秒間の音刺激 (6000 Hz 純音) 呈示後に 4 µl のスクロース溶液が呈示されるため, 音刺激によって報酬の到来が予測できる. C. 非随伴群の実験手続きの模式図. 音刺激と報酬がランダムかつ独立に呈示されるため, 音刺激が報酬の到来に関する情報をもたらさない. D. 瞳孔計測の流れ. 左のパネルは課題中のマウス頭部の動画であり, 中央のパネルは DeepLabCut によって瞳孔の縁の 8 点をトラッキングした動画を示す. 右のパネルはトラッキングした点に対して楕円当てはめを行った動画を示す.

バウト推定

リッキングが瞳孔サイズに与える影響を検討するために, リッキングのバウト・休止パターンを推定した. バウト・休止パターンは反応の一過性のバーストとそれに続く休止期間によって特徴づけられる. バウト・休止パターンは反応間間隔 (inter response time; IRT) が, 2 つの指数分布の混合分布, $P(IRT = \tau) = pe^{-\omega\tau} + (1 - p)e^{-b\tau}$, に従うものとして記述される (Killeen et al., 2002). ω, b, そして p はフリーパラメータであり, それぞれバウト内反応率とバウト間隔, そしてバウト長を表す. 本実験では個体, セッションごとのマウスのリッキングの IRT に上記の 2 つの指数分布の混合分布を当てはめて, 3 つのパラメータを推定した. 推定したパラメータの下で, 各 IRT の尤度を算出し, 反応をバウト内反応とバウトの初発のいずれかに分類した. バウトの初発反応と分類された反応を起点に, 瞳孔サイズの時間を変化の動態を調べることで, リッキングが瞳孔サイズに与える影響を検証する. モデルの当てはめには Julia 言語の確率的プログラミングフレームワークの Turing (Ge et al., 2018) を使用して, モデルの構築と MCMC によるパラメータ推定を行った.

TD 学習

課題中のマウスの内的な報酬予測と報酬予測誤差の動態を明らかにするために, マウスのリッキングデータに対して TD 学習モデル (式 1.1) を当てはめた.

$$V(s_t) \leftarrow V(s_t) + \alpha \cdot \delta \qquad 1.1$$

$V(s_t)$ は状態 s において予測される報酬の価値を表し, α は学習速度を決定するパラメータである. δ は報酬予測誤差を表し式 1.2 によって定義される.

$$\delta = r_t + r_{t+1} + \gamma \cdot V(s_{t+1}) - V(s_t) \qquad 1.2$$

r_t, r_{t+1} はある時点とその次の時点における即時報酬を表し, γ は割引率を表す. 通常の TD 学習モデルでは次の時点の報酬を予測誤差の計算に含めることはないが, モデル当てはめのための前処理の都合により, 上記のように定式化した. マウスの

リッキング回数をモデル化するため, 式 1.1 の報酬予測に基づいて反応率を決定するものとした.

$$A(s_t) \sim Poisson\left(\theta \cdot V(s_t)\right) \quad 1.3$$

$A(s)$ はある時点の状態 s におけるリッキング回数であり, それは報酬予測 $V(s_t)$ と反応率のスケーリングパラメータ θ の積を期待値とするポアソン分布に従うとした. しかし, 実験内では報酬の呈示によって, UR が生じるため, 報酬呈示時の反応率を $A(US) \sim Poisson(\theta)$ とした. さらに, UR は報酬呈示の 2 秒後まで続くため, 報酬呈示後の反応については, 報酬予測ではなく個別のパラメータによってモデル化した, $A(US_{t+1}) \sim Poisson(a \cdot \theta)$. さらに本実験では, セッションごとにパラメータを推定するため, 既にマウスが報酬予測を形成しているものとして, 報酬予測の初期値をフリーパラメータとした. 状態の定義は CS の有無によって決定し, CS が呈示されている期間を状態 1 として, CS が呈示されてない期間を状態 2 とした. 従って本モデルは 2 つの状態における報酬予測の初期値, 学習率 α, 割引率 γ, 反応数のスケーリングパラメータ θ, US 呈示後の反応数を決定するパラメータ a の 5 つをフリーパラメータとして持つ.

データにモデルを当てはめるために, 時系列データを 1 秒のビンに区切り, それぞれのビン内でのリッキング回数, CS の有無, そして報酬の有無を算出した. 随伴群では 1 秒間の CS の呈示直後に報酬が呈示されるため, 上記の処理では CS と報酬が対呈示されないこととなるため, ある時点 t における報酬予測誤差の計算に, 次の時点 $t + 1$ の報酬を含めた. ビン分けされたデータに対して TD 学習モデルを当てはめた. モデルの当てはめには Julia 言語の確率的プログラミングフレームワークの Turing (Ge et al., 2018) を使用して, モデルの構築と MCMC によるパラメータ推定を行った. 推定したパラメータをモデルに与えて, 実データと同じ CS と報酬の呈示の系列データによってモデルを訓練することで, 課題中のマウスの報酬予測と予測誤差の動態を再現した.

結果

パブロフ型条件づけの訓練の最終 3 日間のデータを解析に使用して, CS と US 呈示に伴うリッキングと瞳孔サイズの動態を調べた. 図 1.1.2. に CS と US の呈示時点を 0 秒として, その前後 3 秒間のリッキング数と瞳孔サイズの動態を示した. 随伴群では, CS 呈示によって, リッキング数と瞳孔サイズが上昇していた (図 1.1.2. Aa, Ba の左図). さらにリッキング数と瞳孔サイズは US 呈示直後にピークとなり, 減少へと転じた (図 1.1.2. Ab, Bb の左図). 非随伴群では, CS 呈示によるリッキングと瞳孔サイズの上昇は観察されなかったが (図 1.1.2. Aa, Ba の右図), US 呈示直後にリッキングと瞳孔サイズが上昇した (図 1.1.2. Ab, Bb の右図).

反応が瞳孔サイズに与える影響を検討するため, バウトの開始位置前後でのリッキングと瞳孔サイズの動態を調べた. リッキングのバウト開始の前後 3 秒間のデータを抽出し, その中から CS や US 呈示といったイベントを含むデータを除外して, 解析区間に反応のみを含むデータを使用した (図 1.1.3. A). 群に関わらず, リッキングと瞳孔サイズはバウトの開始から上昇した (図 1.1.3. B, C).

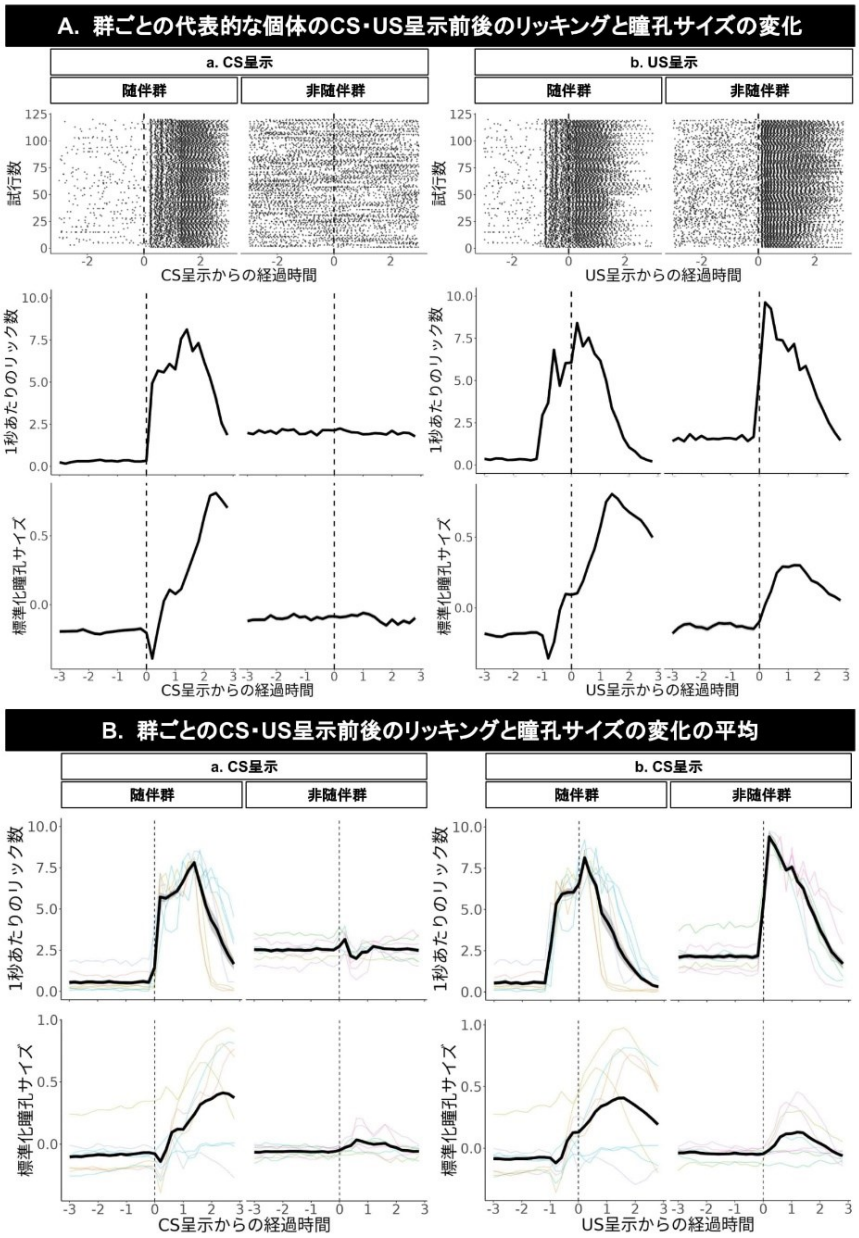

図 1.1.2. 実験 1 の結果.

A. 上から順に随伴群と非随伴群の代表的な個体の, リッキングのラスタープロット, 全試行の平均による 1 秒あたりのリッキング数, そして瞳孔径の変化を示した. B. 群ごとに 1 秒当たりのリッキング数, 瞳孔サイズの時間変化の平均値を示した. A, B ともに CS と US 呈示の前後 3 秒間のデータを示している. 薄く色のついた線は個体データを示す.

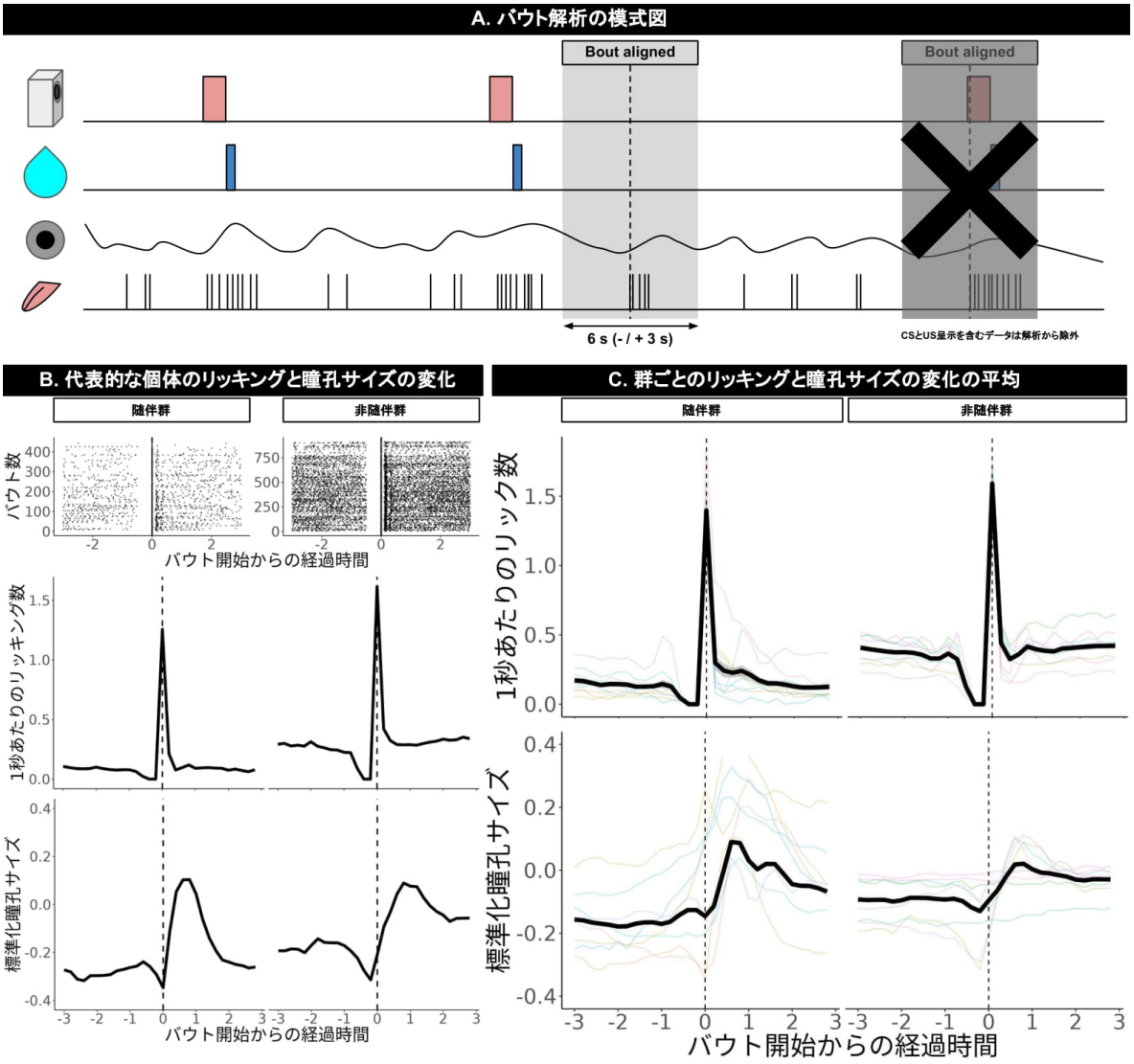

図 1.1.3. リッキングの開始からのリッキング数と瞳孔サイズの時間変化.

A. バウトの解析区間の模式図. リッキングのバウトの開始時点からの前後 3 秒間のデータを使用した. 瞳孔サイズに対するリッキングの影響を検討するために, 解析区間に CS と US 呈示が含まれるデータを除外した. B. 各群の代表的な個体のバウト開始前後 3 秒のリッキング数と瞳孔サイズの変化を示す. C. バウト開始前後のリッキング数と瞳孔サイズの変化の群平均を示す. 薄く色のついた線は個体データを示す.

実験 1 考察

実験 1 では CS によって報酬の到来が予測可能な随伴群と, 予測不可能な非随伴群でリッキングと瞳孔径の動態を比較した. 随伴群では, CS に対するリッキング数の上昇と瞳孔サイズの拡大が観察されたものの, 非随伴群では CS に対する反応は観察されなかった (図 1.1.2. Aa, Ba の右図). このように CS によって報酬が予測されるときにのみ, リッキングの上昇と瞳孔サイズの拡大が生じたことから, リッキングと瞳孔サイズは報酬が到来することへの予測を反映したものと考えられる.

しかし, これらの結果から瞳孔サイズが報酬予測を反映していると結論付けることはできない. リッキングのバウト開始からの瞳孔サイズの動態を検証したところ, 瞳孔サイズはリッキングの開始から拡大することが示された (図 1.1.3.). 既存の研究でもリッキングや身体運動によって瞳孔サイズが拡大することが報告されている (Nelson and Mooney, 2016; Cazettes et al., 2021). 従って瞳孔サイズが報酬予測を反映していることを示すには, リッキングという運動の要因を排除する必要がある.

実験 2：運動表出を抑制しても瞳孔サイズは報酬予測的に拡大した

方法

薬理処置

実験 1 のパブロフ型条件づけの訓練に続いて薬理操作を行った. ここでは 3 日間で 1 ブロックとして全ての個体について計 6 ブロックを実施した. ブロックの初日に, 全ての個体に, 実験開始の 15 分前に生理食塩水を腹腔内に投与した. 2 日目は実験開始の 15 分前にハロペリドール (セレネース注 5mg, 住友ファーマ) を 0.1, 0.2, もしくは 0.5 mg/kg を投与した. 2 日目の実験終了後の 1 時間, マウスはホームケージにて, 自由に水を摂取できた. ブロック最終日は, ハロペリドールの効果を次のブロックへと持ち越さないようにするため, 実験は行わなかった. 全ての個

体は各濃度を 2 回ずつ経験し, 個体によって濃度の投与順を変えることで, カウンターバランスをとった. 実験手続きは実験 1 と同様であった.

結果

ハロペリドールの投与による, リッキングと瞳孔の反応の動態への影響を検討した. 生理食塩水投与時の傾向は両群ともに実験 1 と同様の結果であった. 随伴群では, 生理食塩水の投与時には, CS 呈示によってリッキングと瞳孔サイズはともに上昇し (図 1.2.1. Aa の左図, Ba, Bb の上図の実線), US 呈示直後にピークが観察された (図 1.2.1. Ab の左図, Ba, Bb の下図の実線). 非随伴群では, 生理食塩水の投与時には, CS 呈示によるリッキングと瞳孔サイズの上昇は観察されず (図 1.2.1. Aa の右図, Ba, Bb の上図の点線), US に対する反応のみが観察された (図 1.2.1. Ab の右図, Ba, Bb の下図の点線). ハロペリドールの投与によって, 両群ともにリッキング数が大幅に抑制された (図 1.2.1. A, B). 瞳孔サイズへの影響は, 随伴群では US 呈示直後の一過性の上昇が消失したが, CS 呈示によって生じる瞳孔サイズの拡大は生じていた. 非随伴群では, ハロペリドールの投与時には, 代表的な個体では生理食塩水投与との差は観察されなかったが (図 1.2.1. Aa の右図), 平均的には, CS 呈示によって瞳孔サイズの拡大が生じた (図 1.2.1. Bb の上図の点線). 随伴群と非随伴群の CS 非呈示下でのリッキングと瞳孔サイズを比較すると, 生理食塩水投与時には, 非随伴群でのリッキング数と瞳孔サイズが随伴群を上回っていたが, ハロペリドールの投与時には, リッキングが同程度までに抑制された一方で, 瞳孔サイズは依然として随伴群を上回っていた (図 1.2.1. Ba, Bb). CS 呈示によって生じた瞳孔の拡大が, CS の報酬予測可能性によって異なるか検証するために, 群間で CS 呈示前後の 3 秒間のリッキングと瞳孔サイズの上昇量を算出した. リッキングと瞳孔サイズの上昇量は, 全てのハロペリドールの濃度で, 随伴群が非随伴群を一貫して上回っており, リッキングの上昇量はハロペリドールの濃度依存的に減少したのに対して, 瞳孔の拡大量は濃度に関わらず一定水準を保っていた (図 1.2.1. C). 線形混合モデルによって群間での瞳孔の拡大量の差を検証

した. 全ての個体について複数のデータポイントが存在するため, 個体の識別名をランダム切片に割り当て, リッキングの上昇量の効果を加味しても, 群間での瞳孔サイズの差が確認されるかを検討するために, リッキングの上昇量と群, そしてハロペリドールの濃度を固定効果へと割り当てた. ステップワイズ法によってモデル比較を行った結果, リッキングの上昇量と群, 投与用量を固定効果としたモデルが選択されたため, リッキングの上昇を加味しても, 群間で瞳孔の拡大量に差があることが示唆された (群, F (1, 23.195) = 5.4185, p = 0.029; 用量, F (3, 173.005) = 3.1752, p = 0.0256; リッキング, F (1, 184.886) = 4.7037, p = 0.0314).

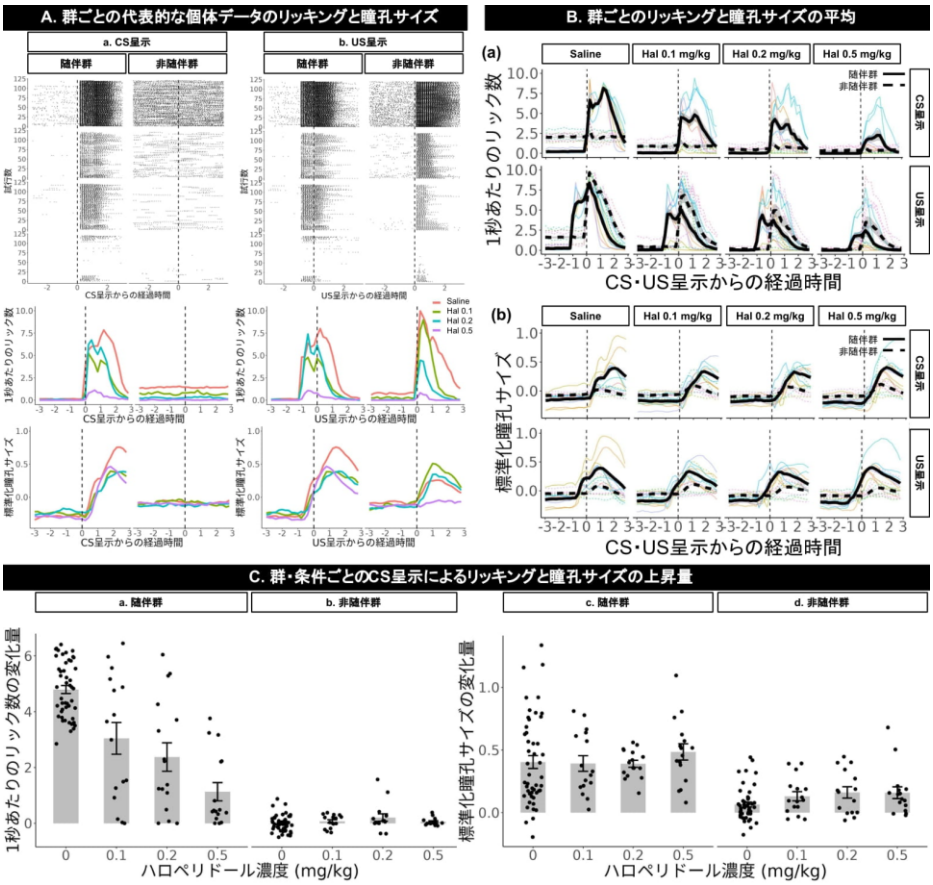

図 1.2.1. 実験 2 の結果.

A. 各群の代表的な個体の用量ごとの, CS と US 呈示前後 3 秒間の, リッキングのラスタープロット, 全試行の平均による 1 秒あたりのリッキング数, そして瞳孔径の変化を示す. 赤, 緑, 青, そして紫はそれぞれ, 生理食塩水, ハロペリドールの濃度, 0.1 0.2, そして 0.5 mg/kg を示す. B. リッキング数と瞳孔サイズの時間変化の群・容量ごと個体間平均を示す. 実線は随伴条件, 破線は非随伴条件を示し, 色のついた薄く示された線は個体データを示す. 左から順に生理食塩水, ハロペリドールの濃度が 0.1, 0.2, そして 0.5 mg/kg の条件を示す. C. 群・用量ごとの CS 呈示に伴うリッキングと瞳孔径の上昇量. 上昇量は個体とセッションごとに, CS 呈示後 3 秒間のリッキング数と瞳孔サイズの平均値と CS 呈示前 3 秒間の平均値の差 (CS 呈示後 – CS 呈示前)によって定義した. エラーバーは標準誤差を示し, 各点は個体・セッションごとのデータの平均値を示す.

報酬予測が瞳孔サイズに影響を与えていることを検証するために, 各個体のセッションごとのリッキングのデータに対して, TD 学習モデルを当てはめて, 実験内での報酬予測と予測誤差の時間的な変化を推定した. モデルによってリッキングの時間的な変化が再現され, CS や US 呈示などのイベントの発生に伴う報酬予測や報酬予測誤差の推定を行った (図 1.2.2.). CS と US 呈示前後のリッキング数の動態は実データとモデルの予測は概ね一致していた (図 1.2.2. Aa, Ba). CS 呈示前後の報酬予測の動態は, CS の呈示によって報酬予測が上昇したが, 非随伴条件では CS による報酬予測の上昇は生じなかった (図 1.2.2. Ab). さらに CS 非呈示下での報酬予測は, 非随伴群では常に一定レベルで形成されていたのに対して, 随伴群ではそれを下回っていた. US 呈示前後の報酬予測の動態は, 両群で CS 呈示前後と同様の傾向であったが, 非随伴群では, US 呈示によって報酬予測誤差が大きく生じていたが, 随伴群ではわずかにしか生じていなかった (図 1.2.2. Bb).

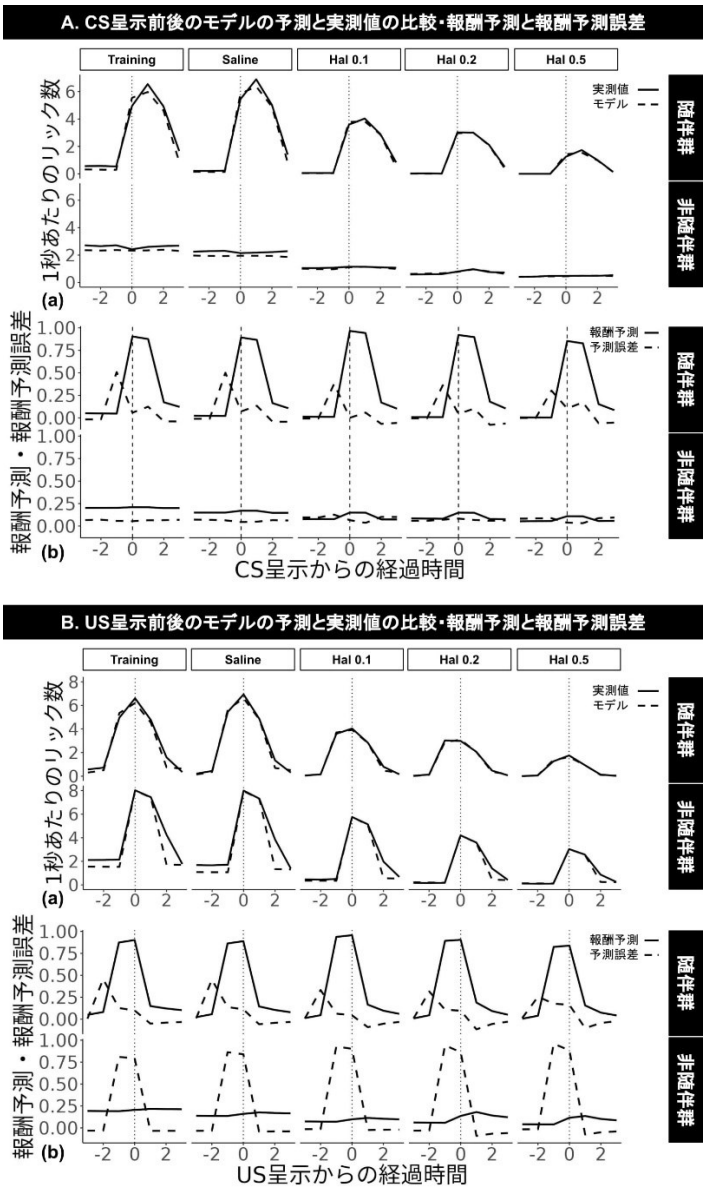

図 1.2.2. TD 学習モデルの当てはめ結果.

A. CS 呈示前後 3 秒間のリッキング数の時間変化とモデルによって予測されたリック数の時間変化 (上) と, モデルによって推定された同時間枠での報酬予測と報酬予測誤差の時間変化を示す. B. US 呈示前後 3 秒間のリッキング数の時間変化とモデルによって予測されたリック数の時間変化 (上) と, モデルによって推定された同時間枠での報酬予測と報酬予測誤差の時間変化を示す. 反応数, 報酬予測, そして報酬予測誤差は 1 秒のビンごとに算出, 推定した.

次に報酬予測によって瞳孔サイズが拡大することを示すために, 推定した報酬予測と報酬予測誤差の系列に CS, US, 各時点におけるリッキング数, と瞳孔サイズの系列を加えて, 1 秒前の各変数から 1 秒後の瞳孔サイズを予測する線形混合モデルを作成した (図 1.2.3. A). 1 秒前の報酬予測, 予測誤差, CS, US, リック数, 瞳孔サイズを固定効果として, 被験体をランダム切片としてモデルの当てはめを行い, ステップワイズ法によってモデルの比較を行った. モデルの比較の結果, 報酬予測, 予測誤差, CS, 瞳孔サイズを固定効果に含むモデルが選択された. 各変数の係数を比較すると, 直前の瞳孔サイズの影響が最も大きく, その次に報酬予測, 予測誤差, そして CS 呈示の効果が大きかった (図 1.2.3. B). この結果から, 報酬予測によって瞳孔サイズが拡大することが示唆された.

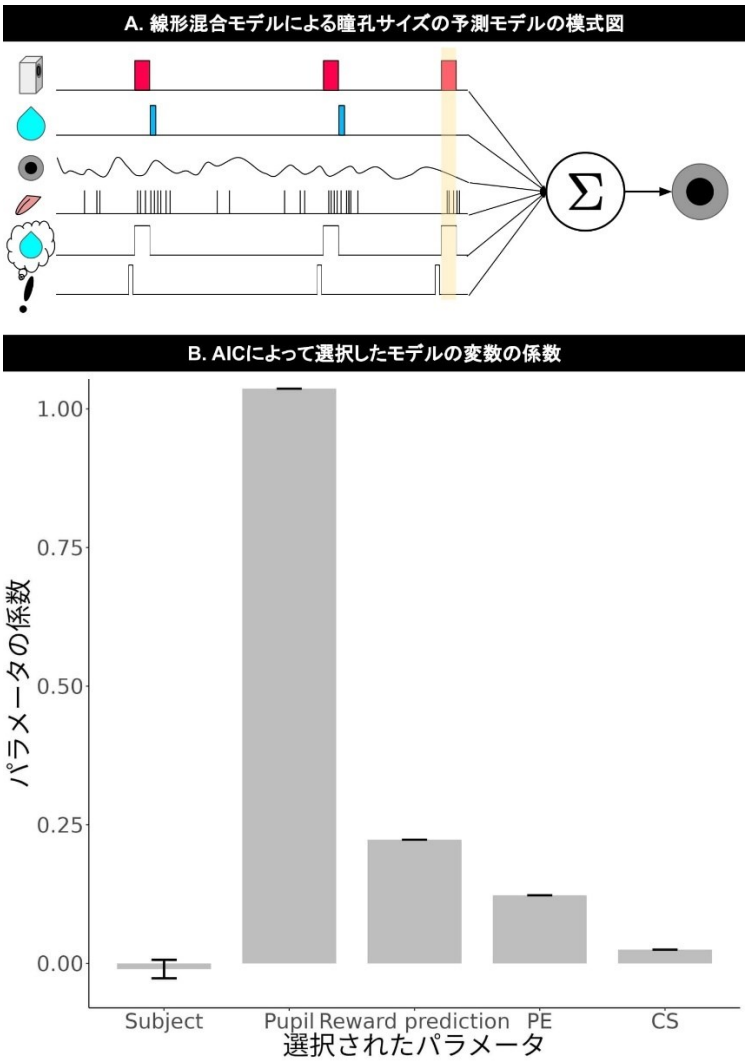

図 1.2.3. 線形混合モデルによる瞳孔サイズの予測.

A. 線形混合モデルによる瞳孔予測モデルの模式図. ある時点における, CS, US, 瞳孔サイズ, リッキング, モデルによって推定された報酬予測, そして報酬予測誤差から, 次の時点における瞳孔サイズを予測するモデルを作成した. B. AIC によるモデル比較の結果, 最も AIC が最小となったモデルの係数を示した. エラーバーは標準誤差を示し, ランダム効果の被験体を除いて同一値であるため, 標準誤差は 0 となっている.

実験 2 考察

実験 2 では, CS 呈示による瞳孔サイズの拡大が, リッキングという運動表出を抑制しても生じるか検討した. ハロペリドールの投与によってマウスのリッキングは, 濃度依存的に抑制されたが, 瞳孔サイズは一定の水準を保っていた (図 1.2.1. Bb). CS による瞳孔の拡大量の差を検討したところ, リッキングの上昇を加味しても, 報酬の予測が可能な随伴条件で拡大量が大きかった (図 1.2.1. C). このことは, 瞳孔サイズが報酬予測によって拡大することを示唆する. さらに強化学習モデルによって, 行動データから課題中のマウスの報酬予測の動態を推定した (図 1.2.2.). 推定した報酬予測と実験的なイベント, マウスのリッキングから, 瞳孔サイズを予測する線形混合モデルを作成して (図 1.2.3. A), 変数選択を行った結果, 報酬予測によって瞳孔サイズが拡大したことが明らかになった (図 1.2.3. B).

随伴群と非随伴群の CS 非呈示下における瞳孔サイズは, 非随伴群が上回っていた (図 1.2.1. Bb). ハロペリドールの投与によってリッキングが抑制されたが, 瞳孔サイズは抑制されなかった (図 1.2.1. B; 図 1.2.1. C). 従ってこの瞳孔サイズの差はリッキングの影響を反映したものではない. 非随伴群では CS 非呈示時にも報酬が呈示されるため, CS の有無に関わらず常にわずかな報酬予測が形成されていたと考えられる. 強化学習モデルによる報酬予測の推定によって, 随伴群では CS に対して高い報酬予測が形成された一方で, 非随伴群では常にわずかな報酬予測が形成されていたことが明らかになった (図 1.2.2. Ab, Bb). 従って, CS 非呈示下の瞳孔サイズの差も, 報酬予測を反映したものである考えられる.

研究 1 総合考察

研究 1 では, パブロフ型条件づけ課題中の瞳孔サイズを計測することで, 瞳孔サイズが報酬予測を反映していることを明らかにした. 実験 1 では随伴群と非随伴群で CS 呈示による報酬の予測可能性を操作して, CS 呈示による瞳孔サイズの拡大を検討した. その結果, 随伴群でのみ CS 呈示によって瞳孔が拡大することが明らかになった. さらにリッキングのバウト開始からの瞳孔サイズの動態を調べ

ることで, 瞳孔サイズがリッキングによって上昇することを示した. 実験 2 では, ハロペリドールを投与することで, リッキングを抑制して, CS 呈示による瞳孔サイズの拡大を検討した. その結果, ハロペリドールによってリッキングが抑制されたにも関わらず, 随伴群では CS 呈示による瞳孔の拡大が観察された. さらに, 強化学習モデルによる課題中の報酬予測の動態を推定して, 瞳孔サイズへの影響を検討した結果, 瞳孔サイズは報酬予測によって拡大したことが示唆された. これらの結果から瞳孔サイズは報酬予測を反映していると考えられる.

瞳孔は交感神経の活性化によって拡大し, 副交感神経の活性化によって縮小する拮抗的な支配を受けている. 瞳孔の交感神経による制御は, 脊髄の頸部及び胸部の中間外側 (IML)のニューロンによって担われている. 副交感神経による制御は Edinger-Westphal 核 (Edinger-Westpharl Nucleus; EWN)のコリン作動性ニューロンによって担われている. 瞳孔径の制御を担う脳領域は複数存在し, 青斑核 (locus coeruleus; LC) と上丘中間層 (superior colliculus; SCi) などが含まれる. LC のニューロンの大部分はノルアドレナリン作動性でありは IML への直接投射によって α1 受容体を介して交感神経の活性化を促す. さらに EWN への直接投射は α2 受容体を介して抑制性に働くことで副交感神経を抑制すると考えられている (Joshi and Gold, 2020). サル, 及びラットで LC のニューロン活動と瞳孔径を同時に測定することで, それらが相関していることが報告されている (Joshi et al., 2016; Liu et al., 2017).

SC は EWN への直接的投射以外にも間接的な経路などを持ち, 一概に瞳孔径への制御の方向性は決定できない (Joshi and Gold, 2020). しかし SCi への微小刺激によって瞳孔径の拡大などが報告されている (Joshi et al., 2016). SCi は主に眼球運動の制御を担っており (Sparks, 1986), 報酬に依存した活動の変化とサッケードの変調を行っていることが報告されている (Ikeda and Hikosaka, 2003, 2007). さらに Redgrave (2010) では視覚刺激に対して DA ニューロンに先だって SCi の

ニューロンが活動することを指摘しており, SC が DA ニューロンの入力元である可能性を指摘している.

研究 1 では瞳孔径が報酬予測的な振る舞いを示すことが明らかになったが, それに関わる神経基盤までは明らかではない. 瞳孔径の制御には複数の脳領域が関与していることからも, 瞳孔径の測定のみでの解釈には限界がある. しかし瞳孔径が報酬予測的な振る舞いを示したことは, Cohen et al. (2012) や Tian et al. (2016) で報酬予測的なニューロンが報告された領域とは異なる領域で予測信号が計算されている可能性を示唆する. 特に瞳孔径を制御している LC や SCi, あるいはその入力元で報酬の予測についての情報が計算されている可能性が考えられる. さらに, 運動抑制のためにハロペリドールを腹腔内投与したが, それでもなお瞳孔径の拡大が生じていることから, 少なくともドーパミン D2 受容体が報酬予測を介した瞳孔径の制御に関与していないことを意味する.

さらに, 先行研究のいくつかの知見は, 従来の報酬予測誤差の計算回路に疑問を投げかけている. 外側手綱核 (lateral habenula; LHb) で報酬予測誤差の符号が反転したものに相当する神経活動が報告されている (Matsumoto and Hikosaka, 2007). さらに LHb のその信号は, 抑制性の GABA 作動性の神経細胞よって構成される RMTg を介して VTA へと間接的に伝えられる (Hong et al., 2011). さらに Tian et al. (2016) によれば, 予測信号源と考えられる領域においても, 報酬予測誤差信号が報告されている. これらの結果は VTA の DA ニューロンの入力元で既に予測誤差が計算されている可能性を示唆する. さらに予測誤差が計算されていることは, さらにその前で報酬予測も計算されていなければならない. これらの知見と瞳孔径が報酬予測を反映したことを踏まえると, 予測信号源としての VTA の GABA ニューロン (Cohen et al., 2012) や Tian et al. (2016) で報告された諸領域とは異なる神経回路が報酬予測誤差の計算に関わっていると考えられる.

研究 1 では検討しきれていない問題が 2 つ挙げられる. 第一に瞳孔サイズが報酬予測によって拡大することは示されたが, 報酬予測の詳細までは本研究では検

討しきれていない. 例えば, 報酬予測には, 報酬量や確率, タイミングなど複数の要因が含まれる (Lowet et al., 2020). 本研究では CS によって確実に報酬が予測される群と, 報酬に関する情報をもたらさない 2 群のみで比較しているため, 報酬の到来と非到来という 2 値予測問題となっており, 瞳孔サイズが報酬量や確率といった連続値的な予測を反映しているものかは定かではない. しかし, 非随伴群で CS 非呈示下の瞳孔サイズが随伴群より高かったこと, そして強化学習モデルによって推定された報酬予測の程度もそれと整合的であることから, 瞳孔サイズが報酬の確率的な予測を反映している可能性はある. 今後は CS の種類によって予測される報酬量や確率を操作して, その時の瞳孔の拡大量を検討することで, 上記の問題が解決されることが期待される. もう 1 つの問題はハロペリドールの瞳孔サイズへの影響である. ハロペリドールは非選択的なドーパミン D2 受容体拮抗薬であり, D3 と D4 などの D2 様受容体のほか, アドレナリン α1 受容体にも結合する. アドレナリン α1 受容体は瞳孔の拡張に関与しており, ハロペリドールはアドレナリン投与による瞳孔の拡張を抑制することが報告されている (Korczyn and Keren, 1980). さらに, LC への電気的刺激はアドレナリン α1 受容体やドーパミン D2 受容体を介して, 中脳のドーパミン細胞の活動が誘発することや, 側坐核のドーパミンの放出を促す (Grenhoff et al., 1993; Park et al., 2017). 瞳孔サイズは LC の活動と強く相関することを考慮すると, ハロペリドールの投与は, LC の活動に変調される中脳や側坐核におけるドーパミンニューロンの活動へと影響を与える可能性がある. これは結果的にハロペリドールが報酬予測や, 報酬予測誤差の計算に影響を及ぼす可能性を示唆している. さらにマウスでは, 身体運動によって聴覚野の活動が抑制されることが報告されており (Nelson et al., 2013), ハロペリドールの投与により身体運動が抑制され, その結果, CS に対する瞳孔の反応性を変調している可能性がある. 実験 2 で, 非随伴群でハロペリドールの投与によって, CS に対する瞳孔の拡大が観察されるようになったのは, その要因を反映している可能性がある. 従って, 今後は, より選択性の高いドーパミン受容体の拮抗薬や, 局所投与

による影響範囲の制限によって, ドーパミンによる瞳孔サイズの影響や, LC を介して行われるドーパミンニューロンへの変調などの影響を詳細に検討する必要がある.

研究 1 では, 深層学習による計測技術によって, パブロフ型条件づけ課題中のマウスの瞳孔の動態を計測し, 強化学習による反応の動態のモデル化を通して, マウスの報酬予測の動態を推定することで, 報酬の予測によって瞳孔サイズが拡大することを明らかにした. 瞳孔のサイズは, TD 学習における予測誤差 (Sutton and Barto, 2018)や, Pearce-Hall モデル (Pearce and Hall, 1980) における刺激への注意といった学習理論との関係も議論されているように (Koenig et al., 2017; Pietrock et al., 2019; Vincent et al., 2019), 刺激への単なる応答してではなく, 学習や予測に関わる感覚運動処理の能動的な処理をしている可能性がある (Ebitz and Moore, 2019). このことを踏まえると, 瞳孔サイズの計測を通して, 動物が自身の学習を促すために環境へと能動的に働きかける可能性が示唆される. このように研究 1 では, 新たな指標の計測や, 強化学習による行動のモデル化を通して, 動物の学習の新たな側面を描き出した.

研究 2：好奇心駆動型強化学習による消去バーストの制御変数の同定

背景と目的

研究 1 では, パブロフ型条件づけによって獲得されたリッキングと報酬予測の動態を捉えることで, 瞳孔の時間的な変化に報酬への予測が反映されていることが明らかになった. こうした短い時間スケールでの行動の変化は, 研究 1 で扱ったパブロフ型条件づけのみならず, オペラント条件づけでも同様に起きていると考えられる. そこで研究 2 では, オペラント反応の瞬間的な動態を強化学習によってモデル化することで, 消去バーストの制御要因の同定を試みる. 消去はオペラント条件づけとパブロフ型条件づけの双方で生じる現象であり, 強化子や US の省略によって, 既に獲得されたオペラント反応や条件反応が減少することを指す. 消去では, 反応は単調に減少すると考えられているが, 消去バーストと呼ばれる一過性の反応増加が生じる場合がある. しかし従来の行動分析学や学習心理学のモデルでは, 反応や反応の強さを決定する媒介変数が単調減少するものとして, 消去中の行動や内部過程をモデル化しているため, 消去バーストのような一過性の反応上昇を説明することはできない (Esber and Haselgrove, 2011; Mackintosh, 1975; Nevin and Grace, 2000; Pearce and Hall, 1980; Pearce et al., 1982; Rescorla and Wagner, 1972). そこで, 強化学習でエージェントに環境の探索を促すために導入された好奇心に着目することで, 消去バーストの説明を試みる. 好奇心の実装として, 予測誤差の累積によるものがあり, これを採用することで, 消去時に一時的な好奇心の上昇が見込まれる. さらに一般的な強化学習モデルと同様に, 過去の強化履歴を反映した行動価値関数をモデルに組み込む. このように, 消去中のオペラント反応の動態を, 単一の過程ではなく, 潜在的な 2 つの過程によってモデル化することで, 消去バーストの

生起を予測し, シミュレーションによって消去バーストの制御変数の同定と, マウスの実験によるモデルの妥当性の検証を行う.

消去バーストは, 強化子によって維持されていたオペラント反応へ, もはや強化子が随伴しなくなった時, その反応が一時的に増加する現象である. 例えば, パソコンのディスプレイに文字が表示されない時に, 何度もキーを連打するような現象である. 応用場面で消去バーストの報告例はあるものの (Lerman and Iwata, 1995, 1996), この素朴な直観に反して, 実験環境でこの現象を支持する研究は殆どない (Katz and Lattal, 2020). Katz and Lattal (2020) によれば, 消去バーストの研究には 2 つの障壁があるとされている. 1 つはバーストの定義の問題である. 消去バーストは, 消去直後の「一時的な」反応率の上昇という定義上, 任意の時間枠で分析をする必要がある. 行動分析学は長期にわたる訓練とセッション全体での反応率のような比較的安定的な行動の特性を扱うが, 短い時間枠では当然のごとく反応率に分散が生じる. そのため, 消去による反応率の上昇を明確に定義することが困難である. もう 1 つの問題は分析の時間枠の問題であり, 最も短いものでは 1 発の反応ごとに分析し, 最長のものでセッション全体での反応率を分析する方法がありうる. 上記のバーストの定義と分析の時間枠が研究間で統一されていないことが, 消去バースト研究の障壁と考えられている (Katz and Lattal, 2020). それに加えて, 消去バーストが生じる理論的な背景が存在しないことも, その原因として考えられる. 行動分析学では行動の予測と制御に目的があり, 現象の制御変数は最重要事項である. 既存の理論やモデルから消去バーストを予測することが可能であれば, その制御要因をシミュレーションなどで明らかにすることができるが, 現在では消去バーストを予測する理論が不在であり, あらゆる強化スケジュールや消去の方法を全て試す以外に方法はない. その一方で, 強化学習で発展してきた好奇心によって探索を促す手法は, この消去バーストが生じることを予測し, その制御変数の特定に役立つ可能性がある.

強化学習における好奇心は, エージェントに探索行動を促し, 環境に関する学習を促進させるために導入された (Schmidhuber, 1990). 特に深層学習を強化学習に取り入れた深層強化学習 (Mnih et al., 2015) の登場により, 膨大な状態空間を扱うことができるようになったことで, 環境内の探索の重要性が増し, 好奇心を利用した強化学習モデルは盛んに研究されている. 好奇心の計算論的な定義には大きく 2 つの種類があり (Pathak et al., 2017), 新奇な状態への探索を促すもの (Bellmare et al., 2016) と不確実性や予測誤差が生じた行動を促すもの (Houthooft et al., 2016; Pathak et al., 2017; Schmidhuber, 1991) があり, これらの双方を満たす形で定義されたものがある (Sekar et al., 2020). この新奇性と不確実性・予測誤差による定義は, 好奇心の異なる側面を表している. 新奇性による好奇心は, エージェントがこれまで観測したことのない状態に対して生じるものであり, 何度も経験することで徐々に減少する特性がある. 不確実性・予測誤差による好奇心は, 何度かの観測を通して徐々に形成される回顧的なものであり, 新奇性とは異なり環境の確率的な挙動に応じて, 過去の経験を通して形成される.

後者の不確実性・予測誤差による回顧的な好奇心の定義は, 消去バーストの生起と関連する. 好奇心は環境の探索を促すために採用されるため, その殆どは環境のモデルに基づく状態予測, つまり次の状態s_{t+1}を現在の状態s_tとそこでの行動a_tの組から予測し, 実際に観測された状態との予測誤差によって好奇心が計算される. この予測誤差は状態のみならず強化子についても同様に計算可能であり, フリーオペラント事態への適用が可能である. さらに, 消去ではベースラインの強化確率に応じて一時的に予測誤差が大きく生じることから, 好奇心を仮定することで, 一時的な反応率の上昇が予測される.

消去バーストは応用場面でしばしば言及されるが (Lerman and Iwata, 1995, 1996), 実験的な研究は少なくその制御変数は明らかになっていない (Katz and Lattal, 2020). そこでフリーオペラント事態の動物の行動を強化学習問題として記述して, 強化学習において探索を促すために導入される, 好奇心という概念をモデ

ルに組み込む. モデルとシミュレーションにより, 消去バーストの発生条件を割り出し, それを基にマウスでの行動実験を行い, 消去バーストの制御変数としての妥当性, 及びモデルの評価を行う.

実験 1：シミュレーションによる消去バーストの制御要因の同定

方法

モデル

提案モデル (Q-learning with hierarchical curiosity module; Q-HCM) は 3 つのモジュールから構成される. それぞれは 1) 行動価値関数の学習を担う Asymmetric Q-learning module (AQM), 2) 予測誤差を基に好奇心信号を生成する Hierarchical curiosity module (HCM), そして 3) 1)と 2) の行動価値関数と好奇心に基づいて反応を生成する Action generator (AG) である. エージェントは環境から与えられた強化子を基に AQM と HCM によって行動価値関数と好奇心を計算し, それらに基づいて AG で反応率を決定する. ここでエージェントの出力する反応は反応間間隔 (inter response time; IRT) とした.

AQM は行動価値関数の推定を行うモジュールで, 予測誤差の符号によって異なる学習率を割り当てた Q-learning モデルである. 予測誤差は式 2.1 によって表される.

$$\delta_t = r_t - Q_t \tag{2.1}$$

ここで r_t は強化子の価値を表し, 強化子が呈示された場合には 1, 呈示されなかった場合には 0 を取る. Q_t は行動価値関数で式 2.2 に従って更新される.

$$Q_{t+1} \leftarrow Q_t + \begin{cases} \alpha^- \delta_t \ \textit{if}\ \delta_t < 0 \\ \alpha^+ \delta_t \ \textit{if}\ \delta_t \geq 0 \end{cases} \tag{2.2}$$

ここで α^- と α^+ は予測誤差がそれぞれの符号の時の学習率を表す. 通常の Q-learning では行動価値関数は強化子の期待値へと収束するが, 異なる学習率では行動価値関数の推定はバイアスがかかる. $\alpha^+ > \alpha^-$ では行動価値関数は期待値より高く推定され, $\alpha^+ < \alpha^-$ では期待値より低く推定される. 多くの研究では $\alpha^+ >$

α^- が採用されており, 実際の動物の行動をより正確に記述できていることから本研究でも同様の仮定を採用する.

HCM は式 2.1 で求めた予測誤差を用いて好奇心を生成する. 好奇心は反応によって生じる予測誤差の予測, つまり予測された不確実性によって定義する. 予測された不確実性は行動価値関数と同様の誤差検出と誤差のフィードバックによって更新される. 従って, まず予測誤差に関する予測誤差 (二次の予測誤差)を式 2.3 によって計算する.

$$v_t = |\delta_t| - \epsilon_t \tag{2.3}$$

δ_t は式 2.1 によって計算された予測誤差であり, その絶対値を使用する. ϵ_t は予測された不確実性, つまり本モデルにおける好奇心を表す. ϵ_t は二次の予測誤差 v_t によって更新し, 式 2.4 によって表される.

$$\epsilon_{t+1} \leftarrow \epsilon_t + \alpha_v v_t \tag{2.4}$$

α_v は予測された不確実性の学習率である. この更新則は通常の Q-learning の強化子を予測された不確実性によって置き換えたものと解釈できる. 予測された不確実性は, 強化確率 $p = 0.5$ で最大となり $p = 0$ 及び $p = 1$ で0 となる, 線対称な逆 U 字型の曲線となる.

AG は式 2.2 と式 2.4 の行動価値関数と好奇心によって反応率を決定する. 反応率を表すパラメータ θ は以下の式 2.5 によって求められる.

$$\theta_t = \frac{1}{a \cdot (1-w) \cdot Q_t + w\epsilon_t} \text{ where } 0 \leq \mathrm{w} < 1 \tag{2.5}$$

a は反応率のスケーリングパラメータであり, w は Q_t と ϵ_t の重みづけパラメータで, 行動価値関数と好奇心を一定の配分で反応率へ反映させる. $w = 0$ で好奇心を一切考慮しないで反応出力を行い, $\mathrm{w} = 1$ では行動価値関数を無視して反応出力を行う. θからの反応への出力は. θをパラメータとする指数分布から IRT を生成するため $p(IRT = \tau) = \theta_t\, e^{-\theta_t \tau}$ となる. まとめると提案モデルは階層的に予測誤差を計算することで強化子の不確実性を近似し, それを好奇心として行動出力の変調を行うモデルである.

シミュレーション

Q-HCM の消去中の反応率の時間的な変化をシミュレートすることで, 消去バーストが生じる環境変数を同定する. Q-HCM では, 反応率は θ によって決定されるため, 任意の環境条件における θ の値を計算する. 好奇心は予測誤差の移動平均によって定義されるため, 一時的に高い予測誤差が生じた時に消去バーストが生じることが予測される. 従ってシミュレーションでは強化確率 p を操作する. まず 200 タイムステップに渡って強化子が確率 p で与えられる条件で θ を計算する. 200 タイムステップ経過後に消去として $p = 0$ で θ を計算する. 消去バーストの定義は前半の 200 タイムステップ目の θ を基準に, 消去中に θ がその値を上回った時点の合計値とした. Q-HCM 上のパラメータの α^+, α^- そして w を系統的に操作して, 各パラメータの組み合わせの下で消去バーストの強度を計算した.

Q-HCM では α^+, α^- そして w の 3 つのパラメータが決まると, 任意の強化確率 p の下でのタイムステップ t における反応率 θ を求めることができる. ある時点 t における予測誤差 δ_t は強化子の有無によって 2 通りの表現が可能であり, 強化子が呈示された場合を $\delta_t^+ = 1 - Q_t$, 呈示されなかった場合を $\delta_t^- = 0 - Q_t$ と表せる. それぞれの予測誤差が生じる可能性は強化確率 p によって決定されるため行動価値関数の更新式は式 2.3 の予測誤差の正負を明示的に表すことで, 式 2.6 のように表すことができる.

$$Q_{t+1} \leftarrow Q_t + \alpha^+ p\delta_t^+ + \alpha^- \cdot (1-p)\delta_t^- \qquad 2.6$$

同様に二次の予測誤差は式 2.7 で表される.

$$v_t = p|\delta_t^+| + (1-p)|\delta_t^-| - \epsilon_t \qquad 2.7$$

と表すことができる. 式 2.7 によって再定義した v_t で式 2.4 の更新式に従って好奇心を更新する. 反応率を表すパラメータ θ は AG における指数分布のパラメータとなっているため, 反応間間隔の期待値を表しているため, 小さい値ほど反応率が高くなる. シミュレーションでは分布からのサンプリングを行わないため, 自然に θ を解釈できるように式 2.8 とする.

$$\theta_t = (1-w)Q_t + w\epsilon_t \qquad 2.8$$

式 2.8 は式 1.5 の逆数からパラメータaを除いたものである. 逆数を取ることで θ の上昇が反応率の上昇として解釈できる. さらに a はスケーリングパラメータであるため, 消去バーストの発生について定性的な影響を及ぼさないため, シミュレーションでは無視する. この式 2.6, 2.7, 2.4, そして 2.8 によって任意の強化確率 p の下での消去バーストの強度を計算する.

結果

モデルのパラメータをそれぞれ$\alpha^+ = 0.4,\ \alpha^- = 0.01$, そして$w = 0.5$ と設定して, 強化確率 p が0.5 の環境の下でのシミュレーション結果を図 2.1.1.に示した. 赤と青の曲線はそれぞれ行動価値関数 Q_t と好奇心 ϵ_t を表しており, 紫の曲線は反応率 θ を表す. 行動価値関数, 好奇心, そして反応率のそれぞれは 200 タイムステップ到達時に漸近値へと収束していた. 200 タイムステップ経過直後から消去が開始しており, 一時的に反応率が上昇して, その後減少へと転じた. 行動価値関数は消去開始直後から単調減少したのに対して, 好奇心は消去開始直後から一時的に増加して, その後に減少へと転じた.

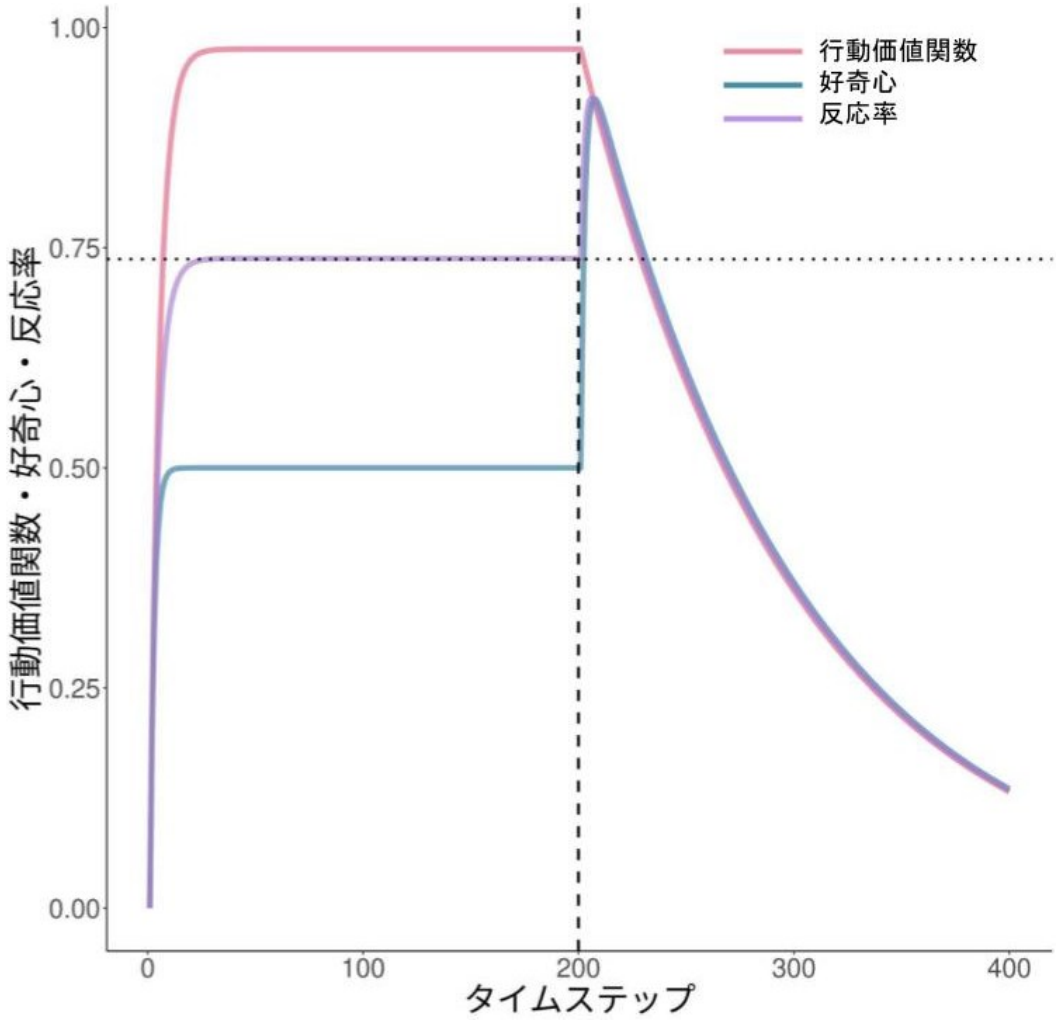

図 2.1.1. Q-HCM によるシミュレーションの結果の例.

前半 200 タイムステップを強化期間, 後半 200 タイムステップを消去とした際の Q-HCM の行動価値関数 (赤), 好奇心 (青), そして反応率 (紫)の時間変化を示す. 破線は消去の開始時点を示す. 点線は強化期間での反応率の漸近値を示す.

消去バーストの強度を, モデルパラメータと強化確率pの組み合わせごとに計算したものをヒートマップに示した (図 2.1.2.). w が上昇するにつれて, 消去バーストが生じる範囲は広くなり, その強度は上昇した. 同様に, 強化確率pが高い条件で消去バーストの強度が上昇した. $p = 1.0$ で $w = 0.5$ (図 2.1.2. 中央右) と $p = 0.5$ で $w = 1.0$ (図 2.1.2.下中央) ではバーストが生じる学習率の組み合わせ, 及びその強度がともに等しかった. 学習率 α^+ と α^- の差が大きいほどバーストの強度が高かった.

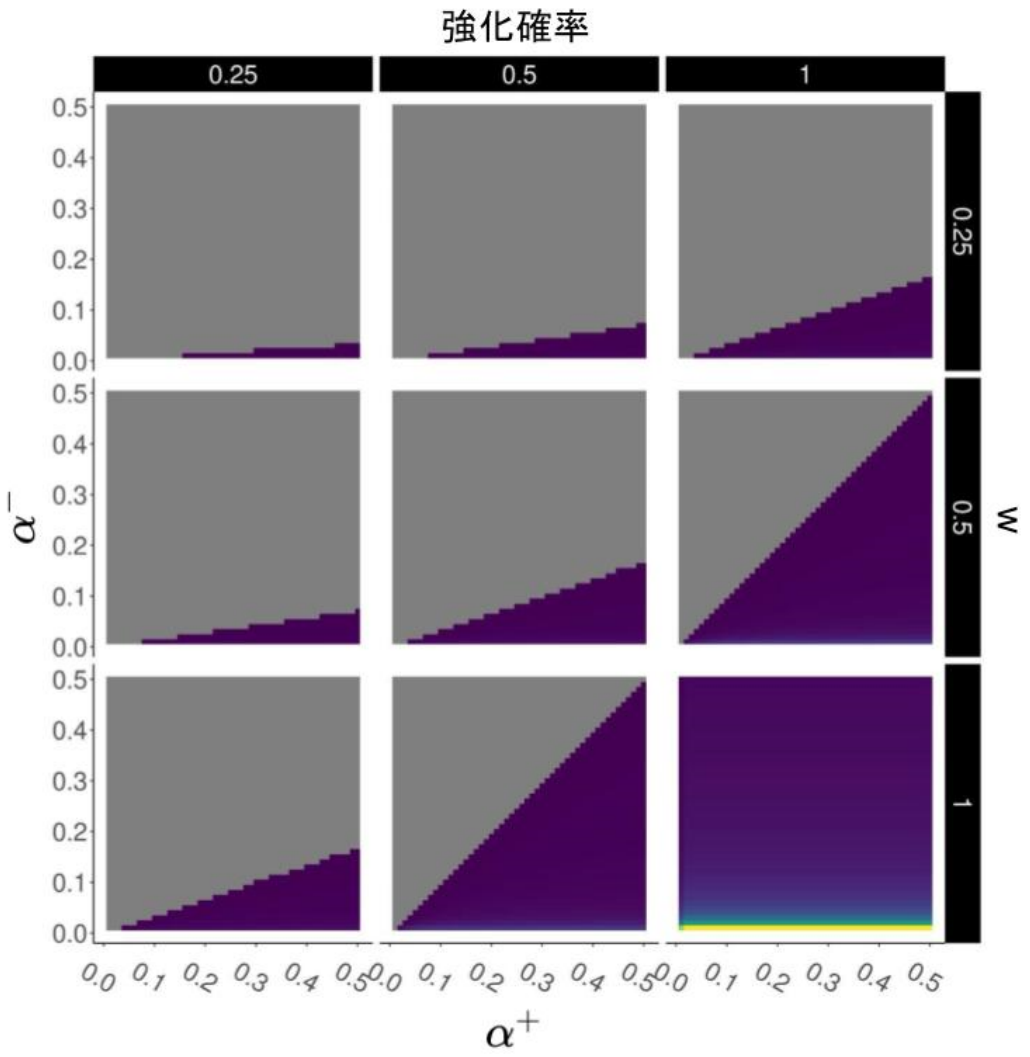

図 2.1.2. モデルパラメータ α^+, α^-, w と強化確率 p ごとの消去バーストの強度.
消去バーストの強度は, 消去時における反応率の上昇分であり, 図 2.1.1.の点線と紫の曲線で囲まれた範囲の面積とした. ヒートマップの灰色で塗りつぶされた範囲は消去バーストが生じなかった条件である. 色が付いている範囲では消去バーストが生じており, 暗い色ほどバーストの強度は小さく, 明るい色に近づくにつれて強度が高いことを示す. ヒートマップの x 軸と y 軸はそれぞれ α^+ と α^- を表している. タイル状に配置されたヒートマップの行によって好奇心への重みづけパラメータ w が異なり, 強化確率pは列によって異なる.

図 2.1.3.は$\alpha^+ = 0.1, \alpha^- = 0.01$, そして $w = 0.5$ における強化期間中の反応率を強化確率 p ごとに表したものである. 反応率は所与のパラメータの下で反応率が漸近値へと収束したときの値を使用した. 反応率は原点を 0 として強化確率の上昇に伴い上昇し, 中程度の強化確率の下でピークとなり, さらなる強化確率の上昇に伴い徐々に減少した.

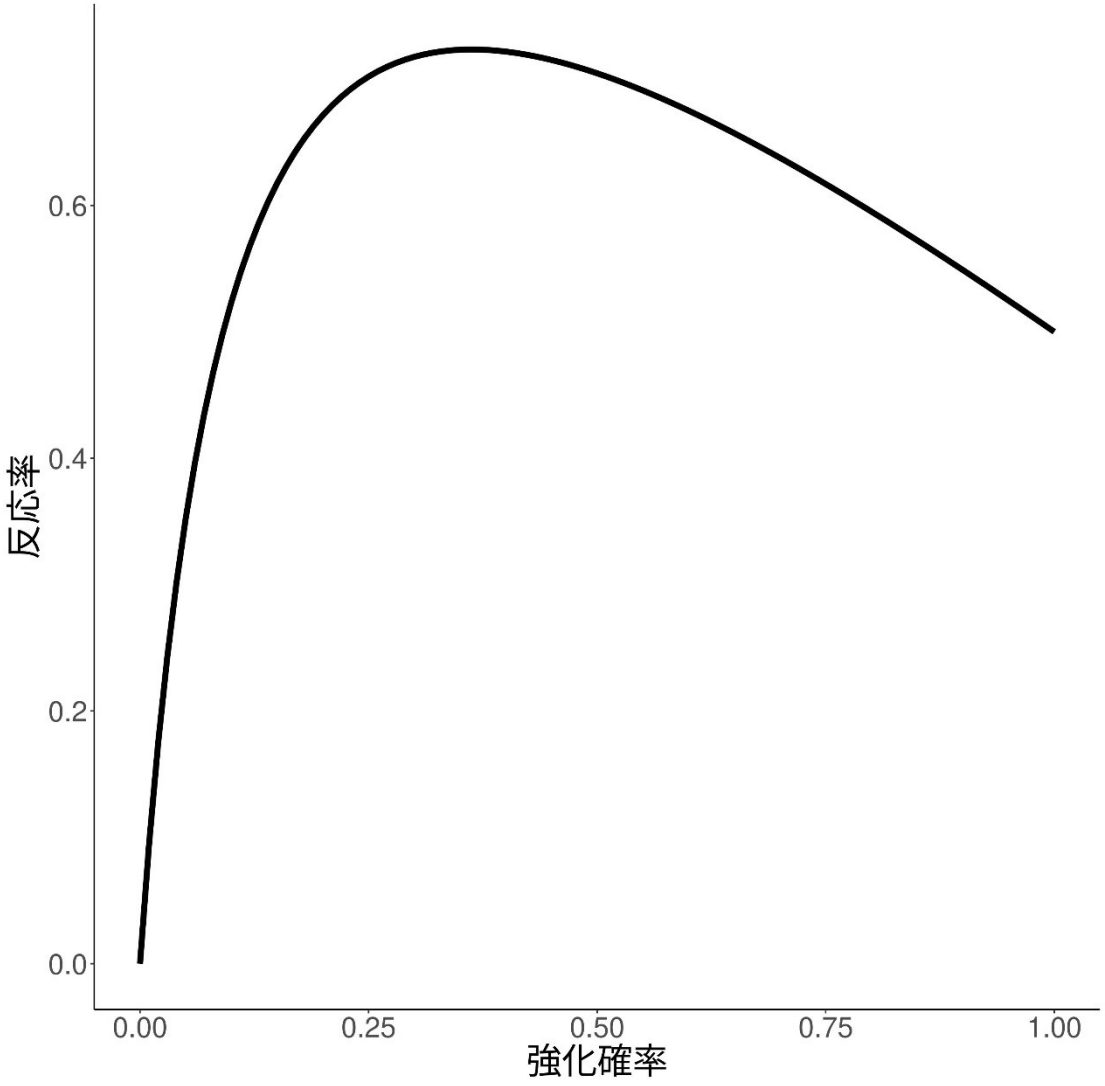

図 2.1.3. Q-HCM から予測される強化期間中の反応率.
$\alpha^+ = 0.1, \alpha^- = 0.01$, そして$w = 0.5$ をパラメータとする Q-HCM によって予測される, 強化期間中の強化確率と反応率の関係を示す.

実験 1 考察

実験 1 ではモデルの挙動を解析的に近似して, 任意の強化確率で訓練した後に消去へ移行する環境を模して, シミュレーションを行うことで, 消去バーストが生じる条件を検討した. 環境側のパラメータである強化確率 p とモデルのパラメータであるα^+, α^-, そして w のそれぞれが消去バーストの生じやすさとその強度に影響を及ぼしていた (図 2.1.2.). 消去バーストについてはシミュレーション結果から, 3 つの結論が導き出せる. 1) 好奇心への重みパラメータ w が大きいほど消去バーストは生じやすい. 2) α^+ と α^- の差が大きいほど消去バーストが生じやすい. 3) 強化確率 p が高い条件で消去バーストが生じやすい. これらの結果はモデルの他のパラメータに関わらず定性的には同じ傾向であった. 従って環境側のパラメータである強化確率が消去バーストを制御している可能性が示唆された.

強化確率が高いほど消去バーストが生じやすいことは, 数少ない消去バーストの実験的知見と一致する. Katz and Lattal (2020) では, 過去の消去を用いた実験をレビューして, FR 1, つまり強化確率が 1.0 では消去バーストの報告があるが, 他の VI や VR といった間歇強化スケジュールでは消去バーストの報告がないことを指摘しており, FR 1 特有の現象である可能性があると提案している. 本実験ではモデルのパラメータを系統的に変化させてシミュレーションしたところ, $p = 1.0$ で最も消去バーストが生じやすく, 強化確率が減少することで消去バーストの生じるパラメータ範囲は縮小し, その強度は小さくなった (図 2.1.2.). モデルの予測ではパラメータに依存するものの強化確率が $p < 1$ の範囲でも消去バーストが生じることが予測されるが, その範囲はかなり限られることが示された. 従って, FR 1 で消去を報告した研究を除いて (Boren, 1961; Keller and Schoenfeld, 1950), 多くの実験では消去バーストが置きづらい実験条件となっていた可能性がある. Q-HCM のシミュレーション結果に基づけば, より強化確率が高い範囲に絞り, 強化確率を系統的に操作して, 消去バーストの生じやすさを検討することで, 強化確率が消去バーストに与える影響が明らかになると期待される.

モデルのパラメータによっても消去バーストの生じやすさ, その強度が変化していた. まず好奇心への重みづけを表す w が大きくなるに従って消去バーストは生じやすくなった (図 2.1.2.). 図 2.1.1. に示されたように消去では行動価値関数は単調現減少するのに対して, 好奇心は一時的に上昇してその後に減少した. 従って消去バーストは好奇心によって生じているものと結論づけることができる. α^+ と α^- 差が大きいほど消去バーストが生じやすかった. 予測誤差の符号によって学習率が異なる場合には学習率の大小関係に応じて行動価値関数にバイアスが生じる. 本実験では $\alpha^+ \geq \alpha^-$ としたため, 行動価値関数は常に強化子の期待値より高く推定された. 行動価値関数が高く推定されることは, 消去に移行した際に生じる予測誤差は通常の Q-learning と比較して高くなる. 従って α^+ と α^- の差が大きいほど行動価値関数は期待値より高く見積もられ, その結果, 消去に移行した際の予測誤差が大きくなり消去バーストが生じやすくなった.

Q-HCM では消去バースト以外にも, 強化確率に対して反応率が上に凸な関数となる予測がなされた. 通常の Q-learning では行動価値関数に対して反応率は単調上昇するため Q-HCM の特徴的な点である. この結果は好奇心によって生じたものである. 好奇心は強化子の不確実性に対する予測であり, 強化子が必ず呈示される, あるいは必ず呈示されない環境で最も小さくなり, 呈示されるか不確実な環境で高くなる. 従って好奇心は中程度の強化確率で高くなり, Q-HCM ではそれによって反応率を制御するため反応率が上に凸な関数となることが予測される. しかし $w = 0$ で Q-HCM は好奇心を無視するため, この予測は w 依存的である. この反応率と強化確率に関する予測は, 既存の研究結果から支持される. Baum (1993) と Baum and Grace (2020) では幅広い強化率の VI スケジュールでハトのキー突きを訓練したところ, 強化率に対して反応率が上に凸な形状となることが報告された. しかし, Baum and Grace (2020) は, 強化子呈示直後の反応を取り除いて解析することで, 反応率は強化率に対して単調上昇することを示した. そして, 強化

子呈示後のキー突きにはフィーダーからキーまで頭を移動させる時間が含まれるため, このような結果になると結論づけている.

実験 2：強化確率はマウスの消去バーストの有無を決定づけた

実験 1 では Q-HCM のシミュレーションによって, 強化確率が消去バーストの制御要因であることが予測された. さらに従来の研究で検討されてきた強化確率の範囲より, かなり限局された範囲でしか生じない可能性がある. 実験 2 では現実のマウスに対して, 頭部固定下でオペラント条件づけを行い, 強化確率を狭い範囲で系統的に操作して, 消去バーストの有無を検証する. さらに Q-HCM では反応率が強化確率に対して上に凸になることが予測された. Baum and Grace (2020) はハトで得られた同様の結果をフィーダーからキーまでの距離という, 実験装置の物理的距離によって生じるものと結論づけたが, 頭部固定下のオペラント条件づけ事態ではマウスは移動することなくリッキング, 及びレバー押しを行うことができるため, Baum and Grace (2020) のような実験装置上の物理的制約は存在しない. 従って, 上記の結果が頭部固定装置でも再現されることは, 強化確率と反応率の関係を規定する異なる要因, Q-HCM によれば好奇心, が存在することを意味する. つまり実験 2 では現実のマウスの実験を通して, 強化確率が消去バーストの制御要因として妥当か, そして強化確率と反応率の上に凸な関数形が再現されるかを検証する. さらに, 瞳孔サイズは不確実性や予測誤差を反映することが報告されており, 予測誤差の累積によって定義される好奇心が瞳孔サイズに反映される可能性が示唆される (Joshi and Gold, 2020; Vincent et al., 2019; Zénon, 2019). そこで, 研究 1 と同様の方法によって瞳孔サイズを計測することで, 消去に伴う瞳孔サイズの動態を調べる.

方法

被験体

被験体として 8 個体のオスの成体マウスを使用した. 全ての個体は実験開始時に過去の実験履歴はなかった. 飼育条件は実験 1 と同様であった.

装置

実験装置はオペラント反応用のレバーを除いて研究 1 と同様であった. レバーはマウスが固定されているプラットフォーム前方に設置されており, 実験開始前にマウスの手がレバー上に乗るように位置を微調整した.

手術

実験 1 と同様であった.

瞳孔計測

実験 1 と同様であった.

手続き

頭部固定下でオペラント条件づけを行った. 実験装置への馴化の手続きは研究 1 と同様であった. 馴化後にレバー押しのシェイピングを行った. シェイピング開始前に, マウスの手がレバー上に乗るように位置を調整した. シェイピングでは毎回のレバー押しに対して強化子が与えられる FR 1 でレバー押しを訓練した. シェイピングはマウスが 200 回の強化子を獲得するようになるまで継続した. 途中で反応をしなくなった個体や, 最初から全く反応しない個体は, シェイピング開始から一時間を経過した時点で中断して, 実験外で追加の給水を行った. 最終的にセッション内で 200 回強化子を得られたところでシェイピングは終了し本実験へと移行した.

シェイピング後に, 強化確率を操作した環境の下でのオペラント反応の訓練へと以降した. 強化確率は 1.0, 0.5, 0.33, そして 0.25 (つまり FR 1, VR 2, VR 3, そして VR 4) の 4 条件であり, 全個体が全ての条件を経験した. まず, いずれかの強

化確率の下で 5 日間の訓練を行った. この期間中は条件で指定された強化確率でレバー押しに対して 200 回の強化子が与えられた. 訓練後にテストへと移行した. テストでは, 訓練時と同様に任意の強化確率でレバー押しが強化されたが, セッション内でランダムに消去期間が 4 回挿入された. 消去期間は 60 秒間レバー押しが自発されなくなるまで継続した. 最後のレバー押しから 60 秒経過後に強化子が一度呈示されて, 訓練と同様の強化期間へと移行した. また消去期間がセッションの開始時や連続しないようにするため, 全セッションを 5 分割して, 最初の区間は強化期間として, 以降の 4 区間内でランダムに消去期間を一度ずつ挿入した. テストでも訓練と同様に 200 回の強化子が与えられた. テストは各条件で 3 回行い, 各テストセッションの翌日には訓練と同様の強化確率で再訓練を行った. 3 度目のテストセッションの後に異なる強化確率での訓練へと移行した.

モデルフィッティング

モデルの当てはめは個体, セッションごとのデータに対して行った. 実データへのモデルの当てはめのために, データから反応間間隔 (IRT) を算出した. 個々の反応に対する報酬の有無も同様に算出して, IRT と報酬の有無の系列に対して, モデルの当てはめを行った. モデルの当てはめでは, Q-HCM と通常の Q-learning model (Vanilla Q-learning model; VQM) のそれぞれを当てはめた. VQM は Q-HCM から好奇心を生成するモジュールを取り除いたモデルであり, 好奇心の学習率 α_v と行動価値関数との重みづけパラメータ w のそれぞれを 0 に設定したものである. それぞれのモデルを当てはめた後に, AIC によるモデルの比較を行った. モデルの当てはめには Julia 言語の確率的プログラミングフレームワークの Turing (Ge et al., 2018) を使用して, モデルの構築と MCMC によるパラメータ推定を行った.

結果

図 2.1.1. は消去期間と強化期間中の反応率の時間変化を強化確率ごとに示したものである. データは消去の開始と強化子の呈示を 0 秒として並べなおしたもの

の全個体・セッションの平均値である. 青線で示される強化期間中の反応率は, 強化確率に関わらず, 一定のレベルで推移していた. 赤線で示される消去期間中の反応率は強化確率に関わらず時間とともに減少していた. 反応が減少する前の目立った反応率の上昇などは確認されなかった.

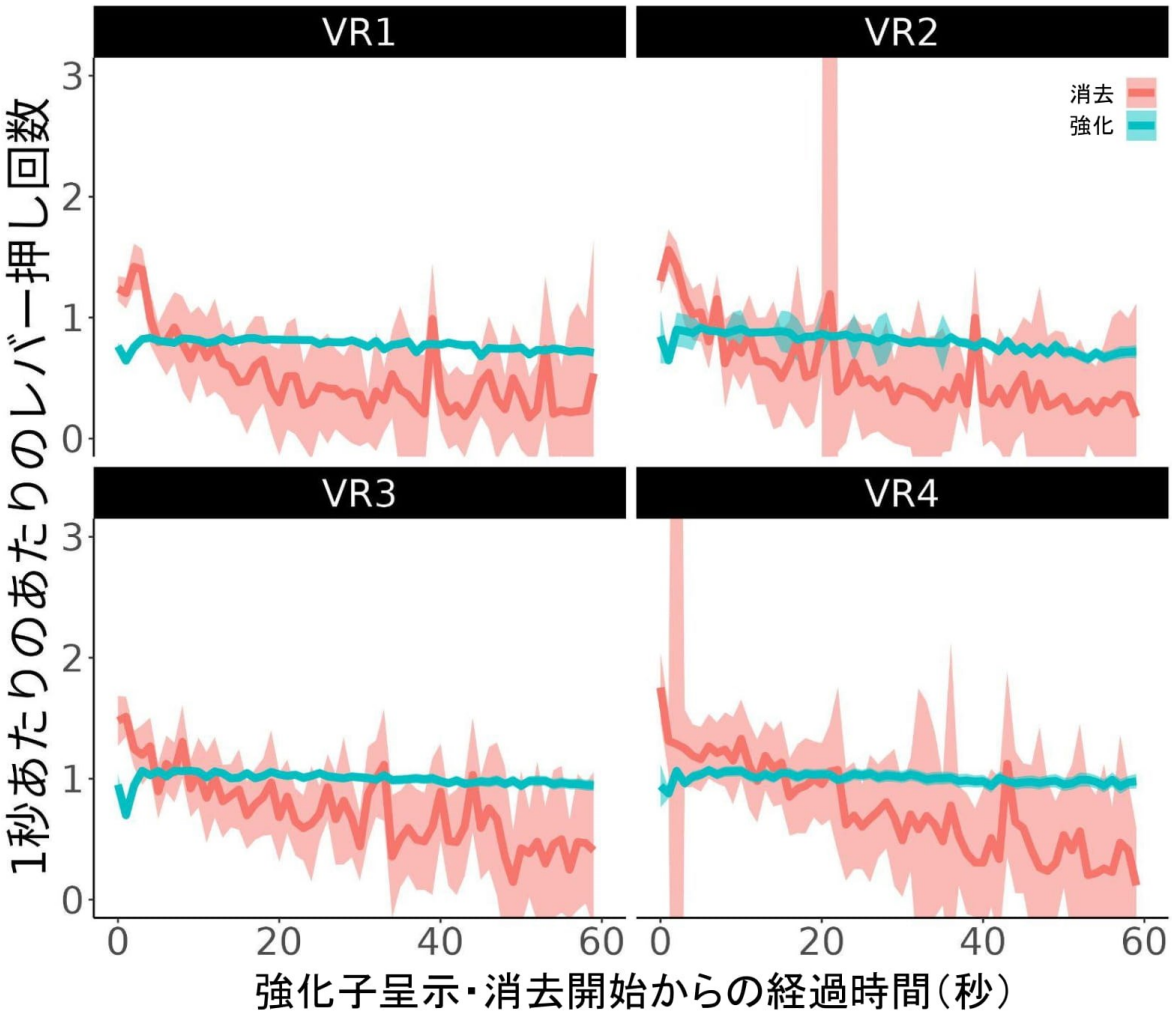

図 2.2.1. 強化期間 (青)と消去期間 (赤)の反応率の時間的変化
強化子の呈示, 消去の開始時点を 0 秒として算出した反応率の時間変化を示す. 薄く塗りつぶされた区間は標準誤差を示す.

図 2.2.2. は図 2.2.1. と同じ時間枠内の反応ごとに, 局所反応率を算出したものである. 局所反応率は IRT の逆数によって算出した. 点線は強化期間中の局所反応率の 95%分位点を示している. 強化確率が 1.0 の条件の消去期間中には, 95%分位点を上回る局所的な反応率が強化期間以上に生じたが, 強化確率が 0.5 以下の条件ではそうした傾向は認められなかった. この 95%分位点を超えた局所反応率をパルスとして以降の解析に使用する. 強化期間中と消去期間中のパルス発生率を比較するために, 強化期間中のパルス発生率に対する消去期間中のパルス発生率の比を条件ごとに算出した. VR 1 の条件でパルス発生率の比が高く, そこから VR 値の上昇に伴い, パルス発生率の比は減少した (図 2.2.3.).

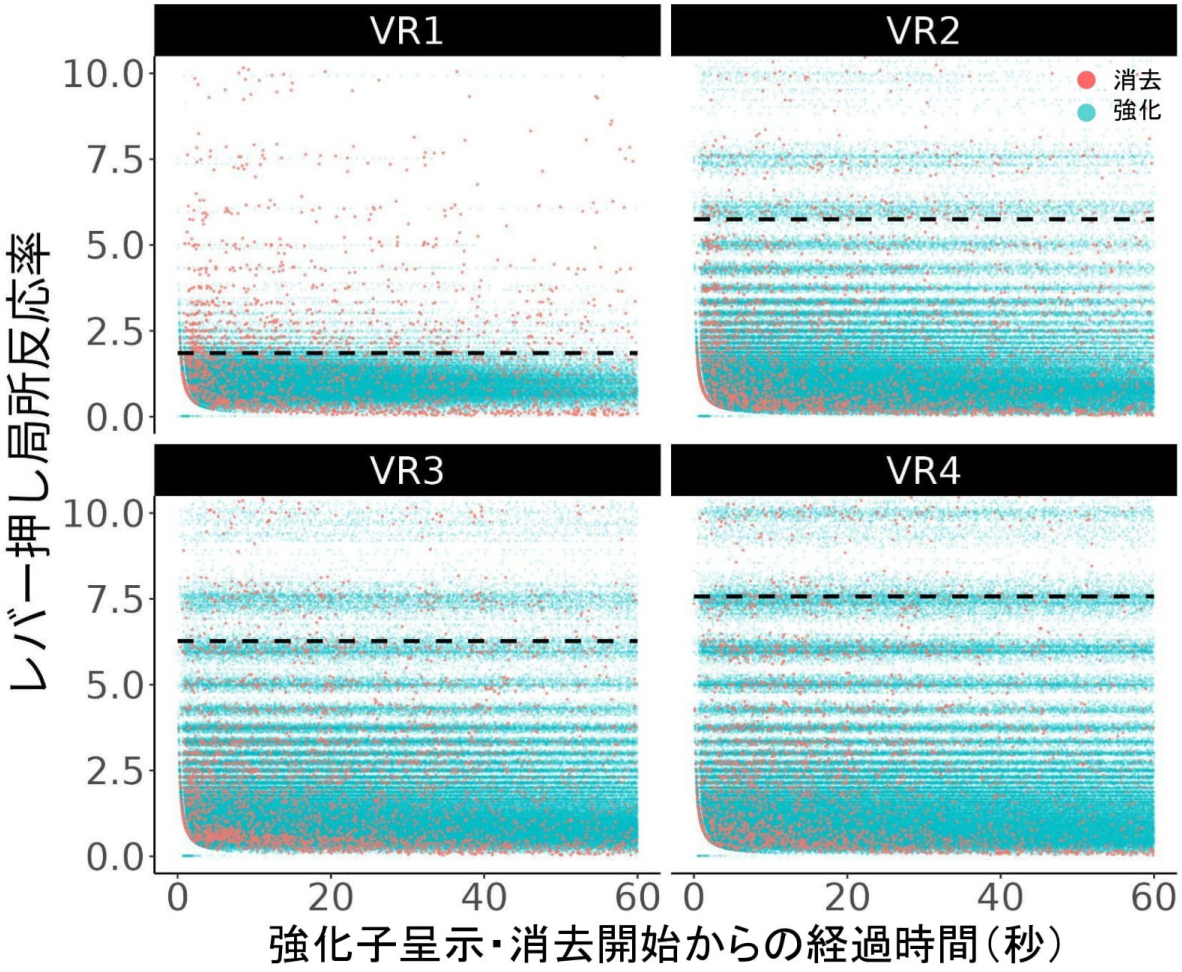

図 2.2.2. 強化期間と消去期間中の各時点において観察された局所反応率.
青で示される点は強化期間, 赤は消去期間中のデータを示す. 局所反応率は IRT の逆数によって算出した. 点線は条件ごとの強化期間中の局所反応率の 95%分位点を示す.

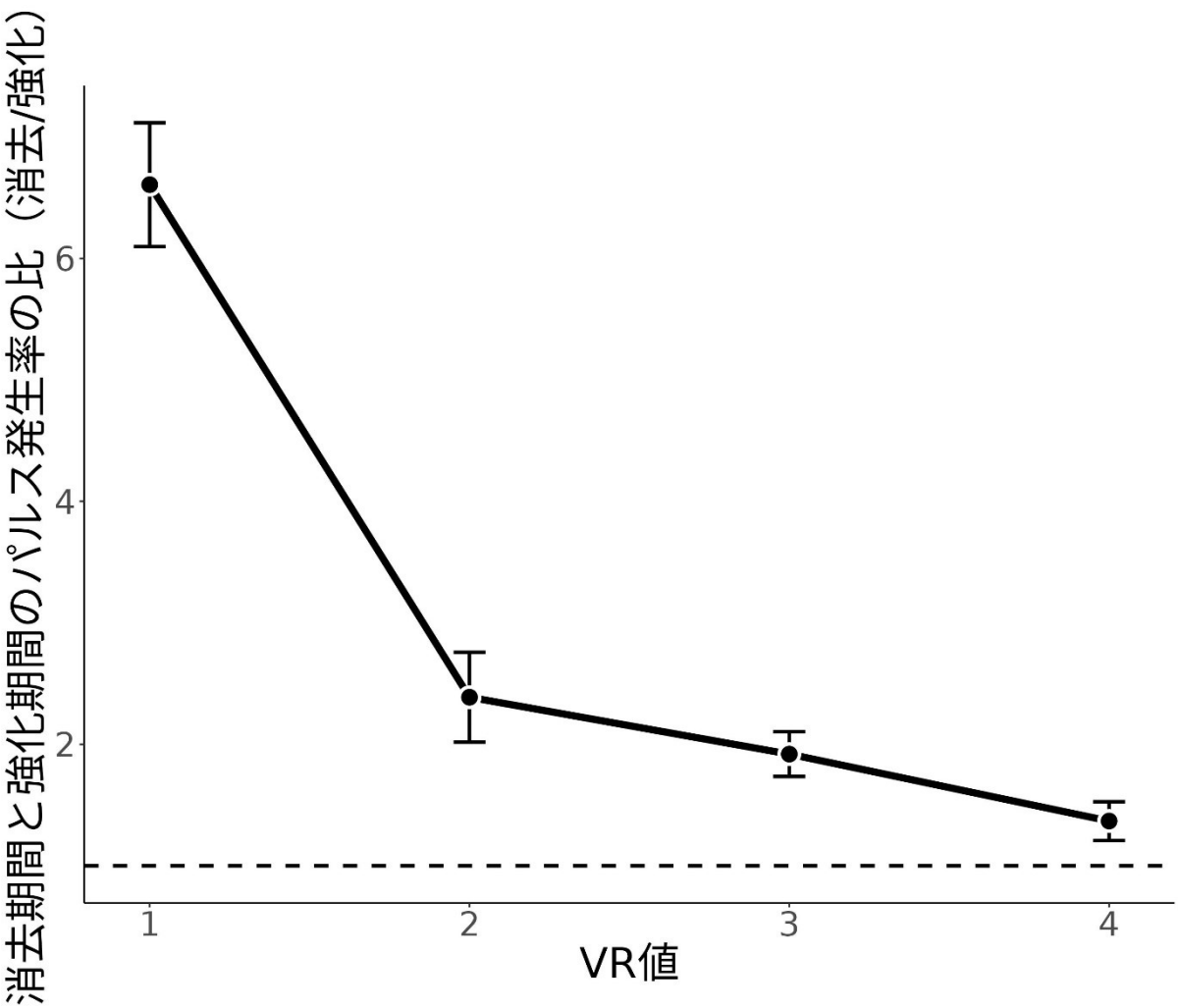

図 2.2.3. 条件ごとのパルス発生率の比

強化期間中のパルス発生率に対する消去期間中のパルス発生率の比 (消去期間 / 強化期間)　を条件ごとに示した. 各点は個体・セッションの平均値であり, エラーバーは標準誤差を示す.

図 2.2.5.は, 各点がパルスが発生した時間を示し, 点を基にカーネル密度推定によってパルスの時間分布を算出したものである. 全ての条件で消去の開始直後にパルス発生率が上昇し, 時間とともに減少していた.

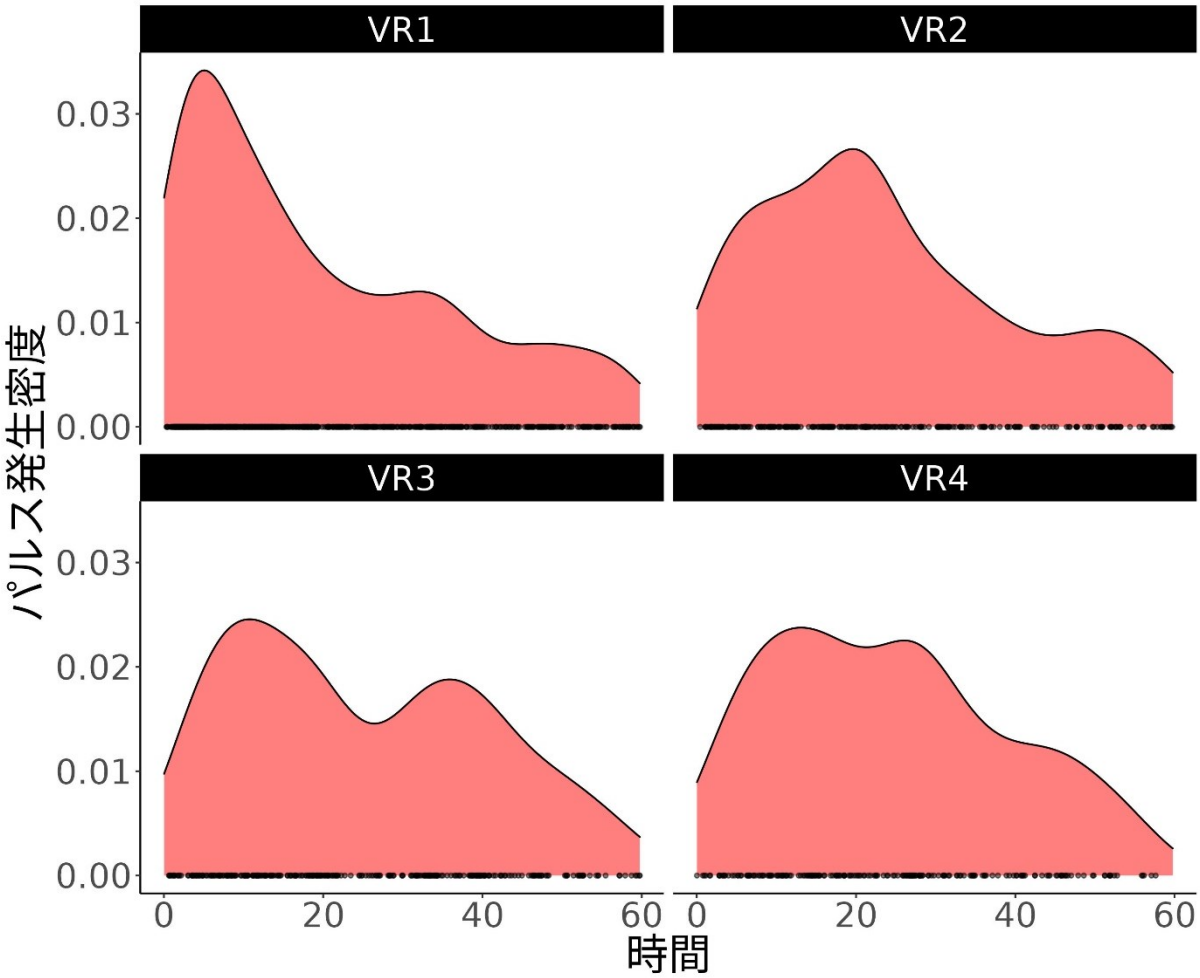

図 2.2.4. パルスの時間分布

全ての個体, セッションのデータをまとめて, 条件ごとにカーネル密度推定によって求めたパルスの時間分布を示す.

図 2.2.6 は強化期間中の反応率を強化確率ごとに示したものである. 強化確率が最も高い 1.0 で反応率が最低となり, 強化確率が減少するに従って反応率は上昇し, VR 3 と VR 4 の反応率はほぼ同程度であった.

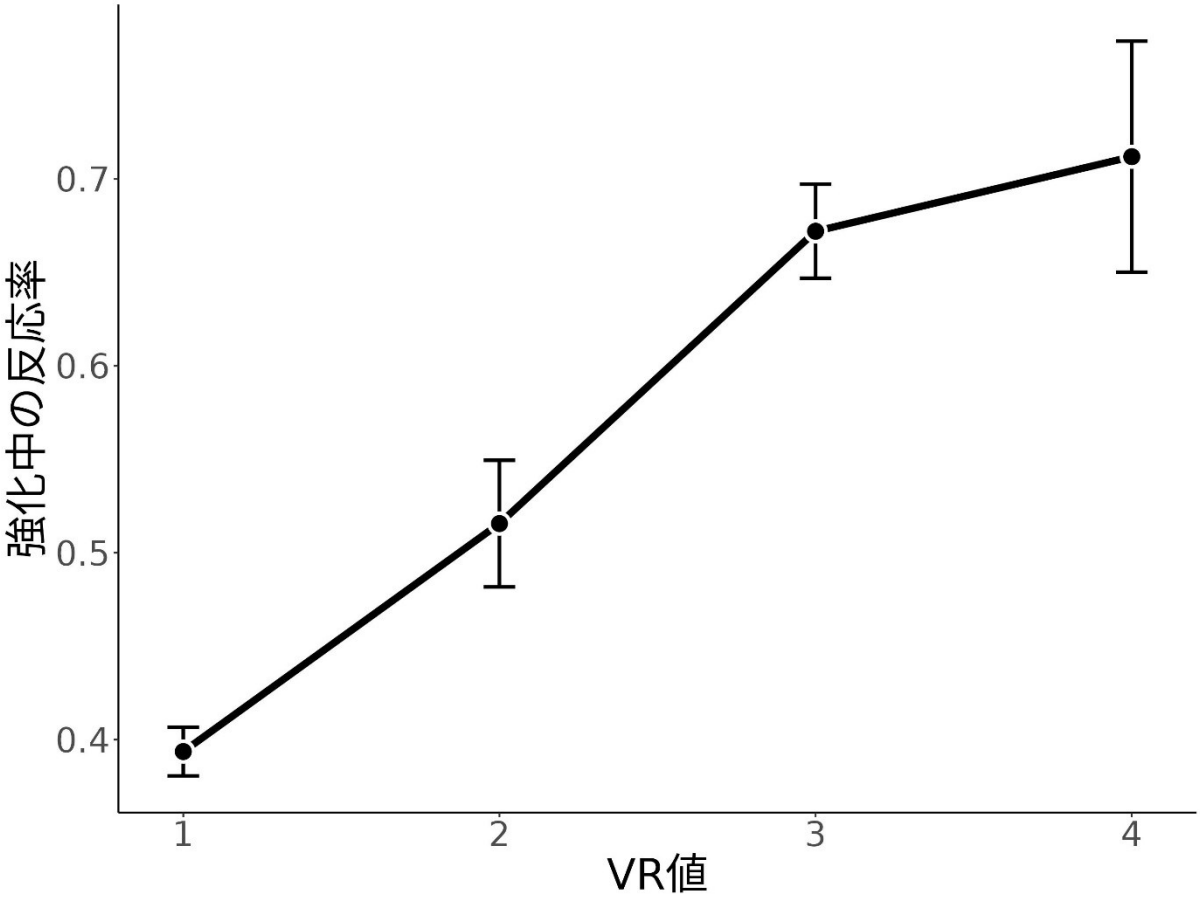

図 2.2.5. 各条件における強化期間中の反応率
条件ごとに消去期間中の反応をデータから除外して反応率を計算した. 各点は個体・セッションの平均値を示し, エラーバーは標準誤差を示す.

Q-HCM と VQM を, 個体とセッションごとのデータに当てはめて, AIC によるモデル比較を行った. AIC によるモデル比較の結果, 概ね Q-HCM の選択割合が高かったものの, 消去バーストが生じた VR 1 条件では他の条件と比較して Q-HCM の選択割合が低かった (図 2.2.6. C). 各個体とセッションごとに AIC によって選択されたモデルと, 実測値の消去前後 60 秒間の反応率の時間変化を比較すると, 両モデルで大まかな傾向は捉えられている一方で, 強化中は実データより過大に, 消去中は過少に反応率が推定された (図 2.2.6. A). モデルによって推定された行動価値関数 Q_t は条件とモデルに関わらず, 消去開始から単調に減少した (図 2.2.6. B 赤線). 好奇心 ϵ_t の動態は Q-HCM では VR 1 条件で消去開始から上昇したが, それ以外の条件では単調減少した (図 2.2.6. B 紫線).

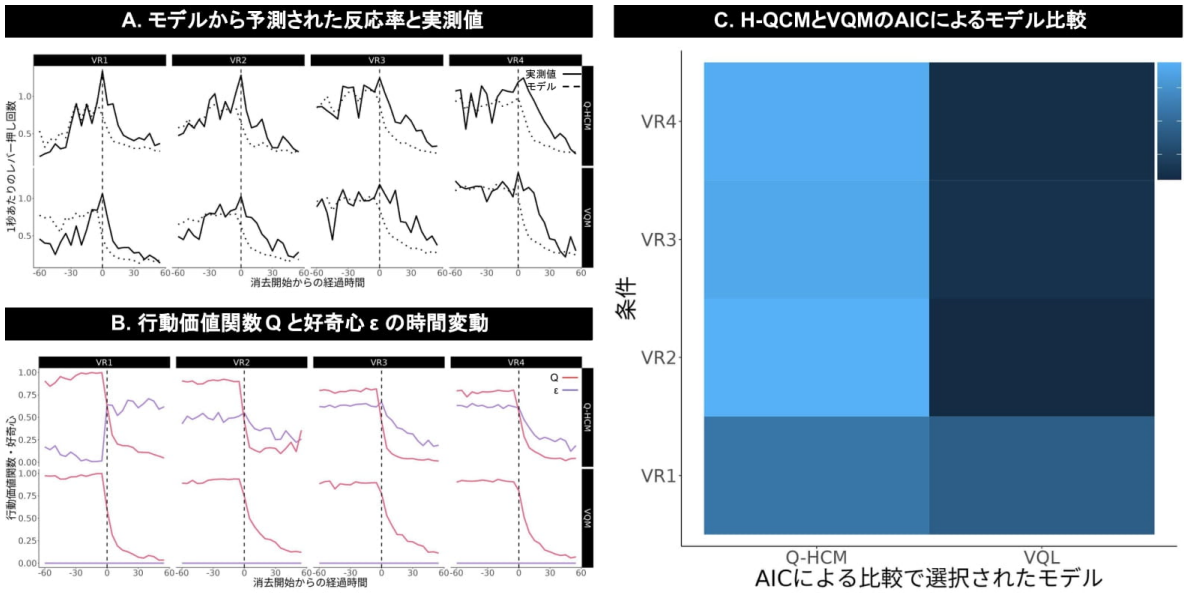

図 2.2.6. Q-HCM と VQM のモデル当てはめ結果.

A. モデルから予測された消去の前後 60 秒間の反応率と実測値の比較. 実線が実測値, 点線がモデルの予測を示し, 個体セッションごとに AIC で選択されたモデルによって, データを分けて表示したため, 上下のパネルには異なるデータが示されている. 上のパネルは Q-HCM による予測, 下のパネルは VQM による予測を示す. B. Q-HCM と VQM の消去前後 3 秒間の内部パラメータ, 行動価値関数 Q と好奇心 ε の時間変化を示す. 赤い線は行動価値関数 Q を示し, 紫の線は好奇心 ε を示す. 上のパネルは Q-HCM, 下のパネルは VQM のデータを示す. C. 条件ごとの AIC によるモデル比較の結果. 色の明るさはモデルの選択された割合を示し, 左が Q-HCM, 右が VQM を示す.

モデルによるバースト強度の予測と実データのバーストの生起しやすさを比較するために, Q-HCM からバースト強度の予測値を算出し, 実データのパルス発生率と比較した. バースト強度の算出は, 個体とセッションごとに, 実験条件と同じ強化確率に設定して, 実験 2-1 のシミュレーションと同様の方法で行った. Q-HCM からのバースト強度とパルス発生率は VR 1 では相関するものの, 他の条件では相関しなかった (図 2.2.7.).

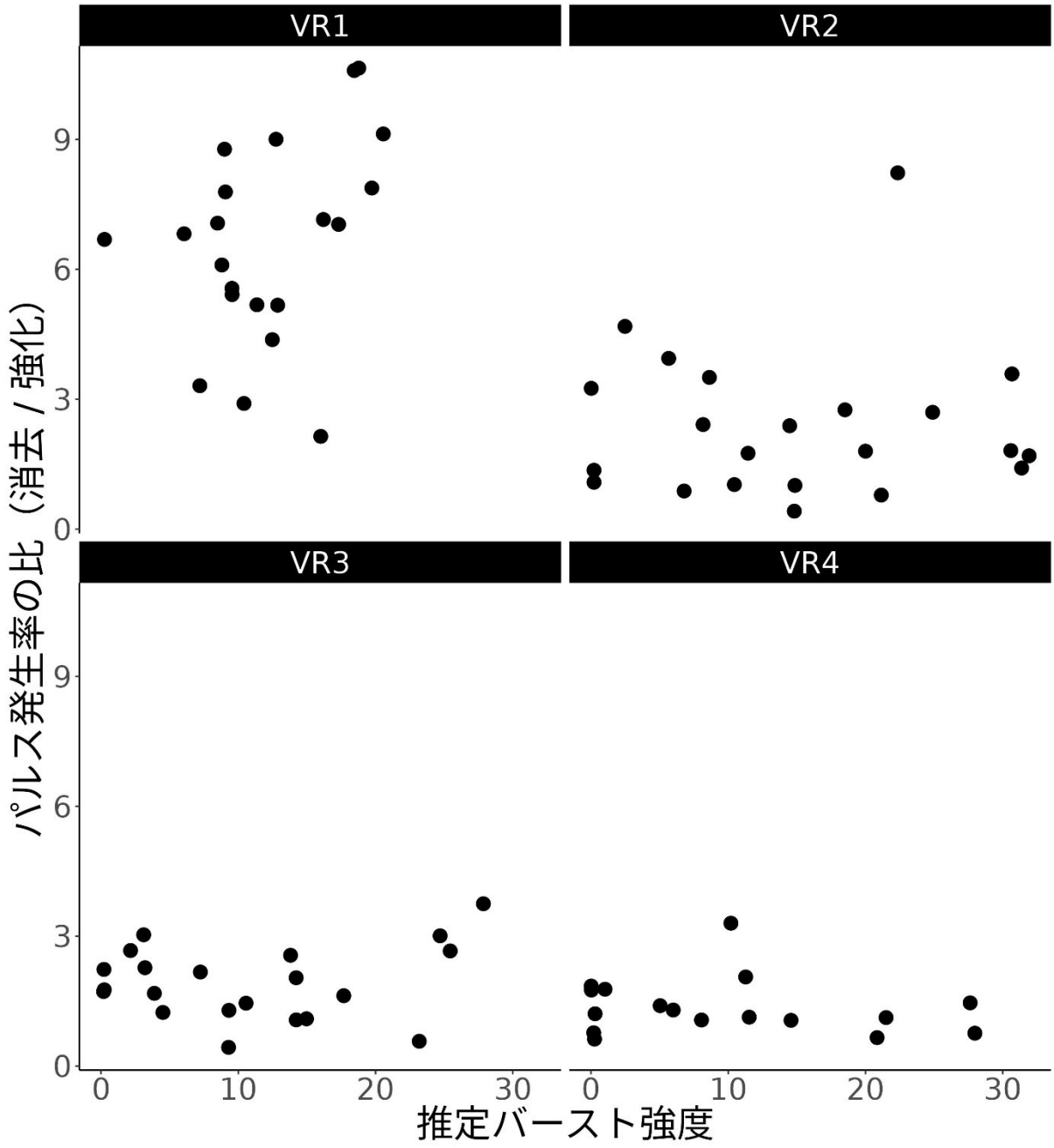

図 2.2.7. モデルから推定されたバースト強度と実際のパルス発生率の比の比較. 横軸はモデルから推定されたバースト強度を示す. バースト強度はモデルの当てはめによって得られたパラメータを用いて, 実験 1 と同様のシミュレーションによって算出した.

瞳孔サイズとモデルにおける好奇心 ϵ_t との対応を検討するために, 消去と報酬呈示前後の瞳孔サイズの時間的変化を図 2.2.8. に示した. Q-HCM によれば VR 1 では消去の開始直後に ϵ_t が増加しているものの (図 2.2.6. B), 他の条件では VR 1 ほどの顕著な増加は生じていなかった. 瞳孔サイズも同様に VR 1 では消去直後に大幅に拡大したものの, 他の条件では VR 1 ほどの拡大は生じなかった.

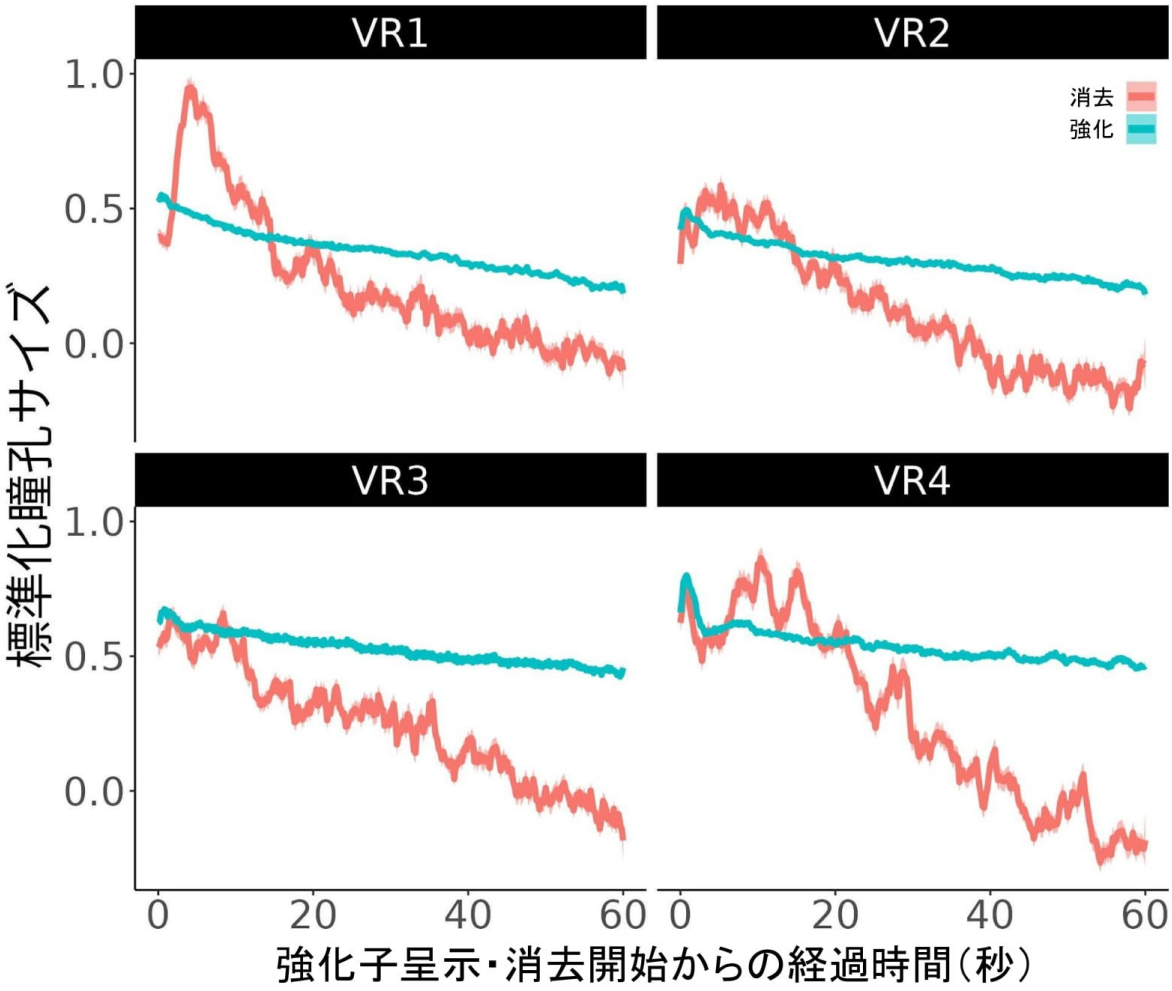

図 2.2.8. 強化期間 (青)と消去期間 (赤)の瞳孔サイズの時間的変化
強化子の呈示, 消去の開始時点を 0 として算出した瞳孔サイズの時間変化を示す. 薄く塗りつぶされた区間は標準誤差を示す.

実験 2 考察

実験 2 では, 強化確率を系統的に操作して消去バーストの有無を検討した. 反応率の時間変化のような, 中程度の時間枠でのデータでは消去バーストは認められなかった (図 2.2.1.). しかし個々の反応の局所的な反応率という, 最小スケールでの解析によって, 強化確率が 1.0 の条件では強化期間を上回る非常に高い局所反応率が消去期間中に観察された (図 2.2.2.). 実データを基に強化期間の 95%分位点の局所反応率をパルスとして解析することで, 強化確率 1.0 で大幅なパルス発生率の上昇が認められた (図 2.2.3.). この結果は Katz and Lattal (2020) の消去バーストが FR 1 特有の現象であるという報告と一致するものである. パルスの発生の時間分布は, 消去の開始に集中しており, 時間とともに減少した (図 2.2.4.). これらの結果を踏まえると, 消去バーストという現象はかなり短い時間枠で, 超高強化率で観察される現象であることが示唆される.

Q-HCM と VQM のそれぞれのモデルを, 実データに当てはめて, モデル比較, 反応率の予測, そして行動価値関数 Q_t と好奇心 ϵ_t の動態を推定した. どの条件でも概ね Q-HCM が選択されたが, モデルに関わらず反応率の予測に実データとの系統的な乖離が生じた (図 2.2.6.). Q_t と ϵ_t の動態は, 前者は消去中に単調減少したが, 後者は VR 1 で顕著に消去開始時に上昇した. しかし, ϵ_t が上昇しているにもかかわらず, Q_t が消去開始直後から急速に減少しているため, モデルから予測された反応率の動態では, 消去バーストが生じていなかった. モデルから予測される消去バーストの強度と実データのパルス発生率の比を比較したところ, VR 1 でのみモデルの予測と実データが相関した (図 2.2.7.). 瞳孔サイズとの動態を検証したところ, VR 1 でのみ, 消去開始直後に一時的な瞳孔サイズの上昇が生じており, モデルの ϵ_t と類似した傾向であった. しかし, モデルと実データに系統的な乖離が生じていることや, 消去バーストが生じている VR 1 で, モデル比較による Q-HCM が選択された割合が低いことから, 現在のモデルではマウスの行動特徴を正確に捉えられておらず, 内部パラメータの推定にも疑問が残る. 今後はモデルの改

善によって, マウスの行動をより正確に捉え, そのモデルを用いて, 瞳孔サイズや消去バーストの生起しやすさを改めて評価する必要がある.

実験 2 では強化確率と反応率の関係も検討した. Q-HCM では強化確率に対して上に凸な関数となることが予測されたが, その傾向は実データでは再現されなかった (図 2.2.5). ただし強化確率の減少に伴って反応率が上昇しており, VR 3 から VR 4 では反応率の上昇がわずかであることから, より低い強化確率では反応率が減少すると考えられる. さらに超高強化率における反応率の減少は Q-HCM 特有の予測であることから, モデルの予測が支持されたと結論付けることができる. Baum and Grace (2020) ではハトで観察された強化率と反応率の同様の関係をフィーダーからキーまでの位置という, 実験装置の物理的制約であると結論づけた. 本実験では頭部固定装置を用いており, マウスの口元に強化子呈示の吸い口があり, マウスの手は殆どレバーを握っていたことから, 上記のような実験装置上の制約は存在しない. それにも関わらず本実験では超高強化率における反応率の減少が再現されたことは, 実験装置上の制約だけでは説明できない.

研究 2 総合考察

研究 2 では, 好奇心駆動型強化学習モデルとシミュレーションによって消去バーストの制御変数を同定し, マウスの頭部固定オペラント条件づけ事態で, その予測を検証した. シミュレーションでは, 強化確率が消去バーストの制御変数であることが予測され, 実験では高い強化確率の下で消去バーストが生じたことから, モデルの予測が支持されたと結論付けられる. しかし強化確率 1.0 では劇的にパルス発生率が上昇していたのに対して, 0.5, 0.33 ではわずかな上昇であった. Katz and Lattal (2020) は過去の研究のレビューによって FR 1 に特異的な現象である可能性を指摘したように, 十分に高い強化確率の下では起きることが示唆されたものの, FR 1 で特に生じやすい現象である可能性がある.

従来の研究で消去バーストが実験的に報告されなかった原因には 3 つあると考えられる. 1 つは強化率の問題である. FR 1 で消去バーストを報告したわずかな研

究を除いて, 消去の研究は本実験で採用したスケジュールより低い強化率を使用している. Katz and Lattal (2020) で FR 1 特有の現象としているが, 本研究のように非常に高い強化率で系統的に消去バーストの有無を検討した研究はない. 本研究では強化確率 0.5, 0.33 でも消去バーストが生じたように, FR 1 ではなくても非常に高い強化率では消去バーストが生じる可能性はある. 本研究では VR スケジュールのみを採用したが, 今後は他のスケジュールでも強化確率を操作することで, その制御変数としての妥当性を検証する必要がある.

2つ目は実験装置の問題である. これまでの行動分析学の実験は自由行動下でのフリーオペラント事態での研究であった. それに対して本研究では頭部固定装置を使用した. この実験装置の違いが結果に影響した可能性がある. 消去では消去誘導性攻撃行動 (Azrin et al., 1966) のように, オペラント反応以外の別の反応が生じることがある. 頭部固定装置では, このような他行動が自由行動下と比較して生じづらい. 過去の研究では消去によって, オペラント反応以外の別の反応が生じたことで消去バーストが観察されづらくなっていた可能性がある.

3つ目は反応のトポグラフィの問題である. オペラント反応にはバウト・休止パターンと呼ばれる時間的な構造が存在する. バウト・休止パターンは強化期間, 消去期間に関わらず観察されるため, 強化期間であっても一時的には非常に高い反応率が観察されることを意味する. すると, 消去へ移行したとしても強化期間での高反応率を超える反応が観察されにくい, あるいは解析上検出することが困難になると考えられる. 応用場面では消去バーストの報告例があるが (Lerman and Iwata, 1995, 1996), そこで対象としている問題行動は, ラットのレバー押しのように連続して何度も繰り返せるような反応ではないため, 消去による上昇が検出されやすかったと考えられる. さらに FR 1 や非常に高い強化率の下ではオペラント反応の一過性のバーストがその途中で強化子の呈示によってたびたび遮られるため, 本研究のような高い強化率で消去バーストが観察されやすい可能性もある. 低反応率分化強 (differential reinforcement of low response rate; DRL) などで, ベ

ースライン反応率を低くなるように訓練することや, 1 回あたりの反応に, 高いコストを要するような反応を採用することでこの仮説を検証することができるだろう.

古い理論では, 消去バーストは攻撃行動 (Keller and Schoenfeld, 1950) によって説明されていた. 一方で本研究では好奇心駆動型強化学習モデルによって消去バーストを説明したが, これは攻撃行動による説明を否定するものではない. 提案モデルでの好奇心は数学的に厳密な定義を持ち, 心的な用語を使わずに言えば環境の不確実性を反映した「何か」でしかない. しかし環境の不確実性によって行動を制御することは一種の情報探求であること, そうした行動が一次的な強化子によって制御されていない, そして強化学習で好奇心という用語が使用されるために, 便宜的に好奇心という用語を使用しているに過ぎない. この環境の不確実性を反映した「何か」はそれ以上でも以下でもない. つまり, ここで好奇心と呼んでいるものを攻撃性と呼ぶこともできる.

背側縫線核 (Dorsal raphe nucleus; DRL) のセロトニンニューロンは, 不確実性を符号化していることが報告されている (Grossman et al., 2022). さらに, DRN のセロトニンニューロンは VTA のドーパミンニューロンに投射しており, その経路を光遺伝子によって選択的に興奮させることで, ドーパミンニューロンは報酬予測誤差と類似した活動を示した (Chang et al., 2021). これは不確実性がある種の強化効果を持つ, あるいは報酬予測誤差に対する変調をしていることを示唆している. その一方でセロトニンは攻撃行動に関与しており (Olivier, 2004), 中枢におけるセロトニン細胞な主要な分布領域である DRN も同様に攻撃行動にも関与する (Bannai et al., 2007; Holschbach et al., 2018). これらは DRN が消去による攻撃行動と好奇心による行動の変調の共通の基盤であると考えられる. 従って提案モデルの予測された不確実性は好奇心としても攻撃性しても双方の解釈が可能である. しかしセロトニンには異なるタイプの受容体が存在し, その分布も異なるため, 攻撃行動と好奇心駆動行動には異なる領域が関与している可能性もがある. 消

去バーストを攻撃行動とみなすか, 好奇心駆動型行動とみなすかは, 今後の神経科学的な研究によって明らかになることが期待される.

好奇心という概念は, 探索を促すものとして強化学習で扱われてきたが, 学習心理学における注意は, 呼称は異なるものの計算論的な定義には共通性が認められる. 学習における注意は Pavlov (1927) で新奇刺激に対する定位反応が「おやなんだ反射」として報告されており, 以降も学習心理学では注意に関わる実験が多くなされてきた (Sutherland and Mackintosh, 2016). 学習における注意の役割が理論的に明示されたのは Mackintosh の注意理論である (Mackintosh, 1975). Mackintosh (1975) では US の予測子に対してより注意が向けられるものとして学習率 α に変動則 $\Delta\alpha_x > 0 \ if \ |\lambda - V_x| < |\lambda - V_y|$ または, $\Delta\alpha_x < 0 \, if \, |\lambda - V_x| \geq |\lambda - V_y|$を与えた. ここで V_x と V_y は異なる刺激 X と Y が獲得した連合強度であり, α_x は刺激 X への注意, つまりは学習率である. ここでは α_x のみを記述したが, 刺激 Y に対する学習率 α_y の変動則も同様に定義される. Mackintosh の注意の変動則とは異なる方法で定義したのが Pearce-Hall モデル (Pearce and Hall, 1980) であり, 予測誤差の大きさによって学習率が変動するものとして $\alpha_{t+1} = |\lambda - \sum V|$ と定義した. ここでは R-W モデルと同じ方法で予測誤差を計算し, その絶対値により学習率が変動する. さらに Pearce et al. (1982) では上記の変動則を修正したものを提案している. Mackintosh (1975) と Pearce-Hall モデル・PKH モデル (Pearce and Hall, 1980; Pearce et al., 1982) は注意の変動則について異なる説明をしており対立的であったが, これらは現在のところ Esber-Haselgrove モデル (2011) によって解決されている. このように, 現在の学習心理学における不確実性・予測誤差による注意の変動は, 強化学習における好奇心との類似性が認められる.

計算論的な定義に類似性は認められるものの, 好奇心と注意の学習への作用の仕方は異なる. 学習理論での注意の変動は明確に学習率 α の変動と対応しており, 強化学習のメタ学習 (Schweighofer and Doya, 2003; Wang et al., 2016) と呼ばれ

る手法に対応する. メタ学習の神経基盤としては特定の神経修飾物質を強化学習におけるパラメータと対応させる解釈がある. 例えばドーパミンを予測誤差として, アセチルコリンを学習率, ノルアドレナリンを逆温度と捉えている (Schweighofer and Doya, 2003). もし注意の変動をメタ学習として捉えるのであれば, アセチルコリンによる海馬の神経可塑性への影響として解釈することができる (Schweighofer and Doya, 2003). その一方で Pavlov (1927) や Kaye and Pearce (1987) で定位反応を注意の指標として扱ったことを踏まえると, 注意の変動は行動の変調を介した環境からの入力の変調という解釈も可能である. この解釈では学習理論における注意の変動 (Esber and Haselgrove, 2011; Mackintosh, 1975; Pearce and Hall, 1980; Pearce, Kaye, and Hall, 1982) はメタ学習というより好奇心駆動型強化学習の一種とみなすことが可能である. このメタ学習と好奇心駆動型強化学習にも計算論的な類似性は認められることから, 全く異なる神経基盤を仮定する必要はないが, 同一のものとみなすべきではないだろう.

研究 2 では, オペラント反応の瞬間的な動態を, 強化学習における好奇心 (Schmidhuber, 1991; Houthooft et al., 2016; Pathak et al., 2017) に注目しモデルを構築することで, 消去中のオペラント反応の瞬間的な動態を捉え, 消去バーストの制御要因を同定した. さらに, 研究 1 と同様に瞳孔サイズを計測することによって, 消去の開始直後に瞳孔サイズが一時的に拡大することが明らかになった. この好奇心と瞳孔サイズの動態から, 動物が消去のような環境の急激な変化に直面したときに, 自身の反応や瞳孔サイズの変調を通して, 能動的に環境からの情報を収集する可能性が示唆された. このように, 研究 2 では, オペラント条件づけと消去バーストを題材に, 動物の学習性の行動の瞬間的な動態を捉えることを通して, 能動的に環境へと働きかける動物の一側面を捉えることに成功した. さらに, 強化学習の知見を取り込んだモデルによって, 行動分析が目的とするところの, 行動の予測と制御, に大きく貢献できる可能性を示すと同時に, 既存の学習心理学におけ

るパブロフ型条件づけとの理論との接点や, 神経科学との接続性も見出すことができた.

研究 3：強化学習によるバウト・休止パターンのシミュレーション

背景と目的

研究 1 と 2 では, パブロフ型条件づけとオペラント条件づけの瞬間的な行動の変化をモデル化してきた. 学習性の行動には, これまで扱ってきたような短い時間スケールで捉えられる変化がある一方で, より長い時間スケールで行動を見た時に立ち現れる性質がある. バウト・休止パターンは, オペラント反応の一過性のバーストと, それに続く長い休止期間によって特徴づけられる時間的な構造であり, 数秒から数十秒といった長い時間スケールで反応を分析することによって, 初めて可視化される. 研究 3 では, このバウト・休止パターンを, 強化学習によってモデル化し, その背後に潜むメカニズムを明らかにする.

動物は日常生活の中でさまざまな反応に従事する. ヒトの場合は, 仕事, 勉強, スポーツ, そしてテレビゲームなどがあり, ネズミの場合は, 毛づくろい, 採餌, 探索, そして捕食者からの逃走などが挙げられる. 具体的な反応は生物種, あるいは個体によって異なるが, それらには共通の行動上の特徴が存在する.

バウト・休止パターンは多くの生物種, 行動に共通して観察される特徴の 1 つである. 動物が行う反応は時間的に一様に分布するのではなく, 短い期間に集中して自発されることが多い. 例えば, オペラント条件づけの実験では, ラットがレバーを短期間に何度も押し, その後レバーを押すのを止める. しばらくすると, またレバーを押し始める. このように, レバーを押す期間と押さない期間が実験中に何度も入れ替わる. このような短時間の反応のバーストと長い休止からなる時間構造は, ヒトのメールや手紙のやりとりや (Barabasi, 2005), 牛の採餌 (Tolkamp and Kyriazakis, 1999), ショウジョウバエの歩行 (Sorribes et al., 2011)など, 様々な種や反応で観察されている.

行動分析学におけるバウト・休止パターンは古くから報告されており (Gilbert, 1958), このような時間的パターンが存在することは, オペラント反応には少なくとも 3 つの成分が存在することを意味する. それはバウト発生率, バウト内反応率, そしてバウト長である. 一方で Shull et al. (2001) がバウト・休止パターンを可視化する手法を開発し, 環境の操作の種類によってバウトを構成する成分に対する影響が異なることが示されるまでは, 反応率という単一の指標でのみ個体の行動を評価していた. Shull et al. (2001) は反応間間隔 (IRT) の相対頻度分布を基に, 生存関数を描き, y 軸を対数変換した対数生存関数上で, 2 本の直線から構成される折れ曲がった曲線が表れることを示した. これは IRT の分布が 2 つの指数分布の混合分布 (bi-exponential model; 二重指数分布) であることを意味しており, Killeen et al. (2002) によって $P(IRT = \tau) = pe^{-\omega\tau} + (1 - p)e^{-b\tau}$ と定式化された. ここで p は 2 つの指数分布の混合割合を表し $\frac{1}{1-p}$ は 1 回のバウトあたりに含まれる平均反応数を表す. ω と b はそれぞれの指数分布のパラメータであり, それぞれバウト内反応率とバウト発生率を表す. いくつかの研究では, この二重指数分布からの逸脱が報告されており, 異なる分布を当てはめることが提案されてはいるものの, 行動を評価する上で, 定性的に大きな違いは生まれないことが明らかになっている (Tanno, 2016).

バウト長, バウト内反応率, バウト開始率は, 動機付けに関わる操作とスケジュールタイプの操作によって影響を受ける (Brackney et al., 2011, 2012; Podlesnik et al., 2006; Shull et al., 2001, Shul et al., 2002; Shull et al., 2004; Shull, 2004; Tanno, 2016). 動機づけの操作としては, 強化率, 反応-強化の随伴性, 遮断化レベルへの操作がある. スケジュールタイプの操作の例としては, 変動間隔 (variable interval; VI) スケジュールに微小な変動比率 (variable ratio; VR) スケジュールを連結スケジュール (tandem) によって追加する方法がある. これらの実験操作とバウト成分は以下のように整理できる.

バウト長は,

- 強化率の上昇によって上昇, もしくは変化しない (Shull et al., 2001; Shull et al. 2004)
- 遮断化レベルの上昇によって上昇, もしくは変化しない (Podlesnik et al., 2006; Shull et al., 2001; Shull, 2004)
- 消去によって減少, もしくは変化しない (Brackney et al., 2011; Cheung et al., 2012; Shull et al., 2002)
- VI スケジュールへの, 連結スケジュールによる微小な VR スケジュールの追加により上昇する (Brackney and Sanabria, 2015; Shull et al., 2001; Shull et al., 2004)

バウト開始率は

- 強化率の上昇によって上昇する (Shull et al., 2001; Shull et al., 2004)
- 遮断化レベルの上昇によって上昇する (Podlesnik et al., 2008; Shull et al., 2001; Shull, 2004)
- 消去によって減少する (Brackney et al., 2011, 2012; Shull et al., 2002)
- VI スケジュールへの, 連結スケジュールによる微小な VR スケジュールの追加により変化しない, もしくはわずかに減少する (Brackney and Sanabria, 2015; Shull et al., 2001)

ここではバウト内反応率については詳細な説明は省略するが, 連結スケジュールによってわずかに上昇するか, 変化しないことが報告されている (Shull et al., 2001; Tanno, 2016; Matsui et al., 2018). このように過去の研究では, 環境操作がバウト成分に与える影響を検討してきた. それを評価するための記述的なモデルも Killeen et al. (2002) 以降にいくつか提案されている (Brackney et al., 2011, 2012; Matsui et al., 2018; Tanno, 2016).

オペラント反応にバウト・休止パターンが観察されることは, 多くの研究で示されてきたが, なぜこうした構造が生じているのかは明らかになっていない. Smith et al. (2014) は, 並立 VI VI の選択事態で, バウト・休止パターンを生む要因として選択とコストの 2 つがある可能性を実験的に示した. ハトのキー突きを単一スケジュールで訓練した場合, バウト・休止パターンが生じないことが報告さ

れている (Bennett et al., 2007; Bowers et al., 2008). そこで Smith et al. (2014) は並立 VI VI 事態で, 選択切り替え後遅延 (changeover delay; COD) の有無を操作した. COD は複数の選択肢が存在する事態で, 選択肢の切り替えてから一定時間は強化子が与えられない手続きである. この COD によって, 個体が急速に左右への反応を切り替えることを防ぐことができる (Herrnstein, 1961). Smith et al. (2014) は, COD がない条件では, 対数生存関数は 1 本の直線によって表されたのに対して, COD がある条件では, 対数生存関数は 2 本の直線から構成される, 折れ曲がりのある曲線となることを報告した. 従ってハトでは明示的に選択肢が存在し, 選択の切り替えにコストが生じる場合にのみ, バウト・休止パターンが生じる. この結果は選択とコストという 2 つのプロセスによってバウト・休止パターンが生じることを示唆する. しかし, 現実の動物を用いた実験では, 実験上では統制不可能な要因や, 個体が内在する他のプロセスの影響を無視することはできないため, 選択とコストという 2 つのプロセスが, バウト・休止パターンを再現するのに十分であるかは定かではない.

そこで実験 1 では, 選択とコストという 2 つの仮定によって, バウト・休止パターンが生じるかをシミュレーションによって検証する. 実験 1 ではこれら双方を組み込んだモデルでバウト・休止パターンが再現できること, どちらか一方を欠損させたモデルでは再現できないことを示すことで, これら 2 つのプロセスがバウト・休止パターンの再現に十分であることを示す. さらに実験 2 では過去の研究で報告されてきた環境操作とバウト構造の関係が再現することでモデルの妥当性を検討する. これらのシミュレーションを通してバウト・休止パターンを生成する機械論的な説明を提供するとともに, その神経基盤を考察する.

実験 1：選択とコストによるバウト・休止パターンの再現

モデル

提案モデルは 3 つの状態からなるマルコフ過程によって動物の行動をモデル化する (図 3.1.1. 上). 3 状態のうち, 2 つの状態は「オペラント」と「他行動」であり, それぞれで, エージェントはオペラント行動か他行動に従事する. 3 つ目の「選択」では, エージェントはオペラント行動と他行動のいずれかを選択する. 「選択」は, 実験中に動物が利用可能な反応から自由に選択している事実を反映している. さらに提案モデルでは, ある反応から別の反応に切り替えることにコストを仮定した. 遷移にコストがかかる場合, 高速な切り替えは最適ではないため, 動物は同じ行動を続けるか, 遷移を行うかを判断しなければならない. 提案モデルを, 選択なしモデルと, コストなしモデルという, 2 つのノックアウトモデルと比較する (図 3.1.1. 中央・下). それぞれのモデルは, 提案モデルから選択とコストの 2 つのプロセスのうちの 1 つを取り除いたものである. 選択なしモデルでは, エージェントは与えられた状況下でオペラント行動のみに従事する. コストなしモデルでは, オペラント行動と他行動を選択できるが, 自己遷移を仮定せず, 一回の反応ごとに選択へと移行する.

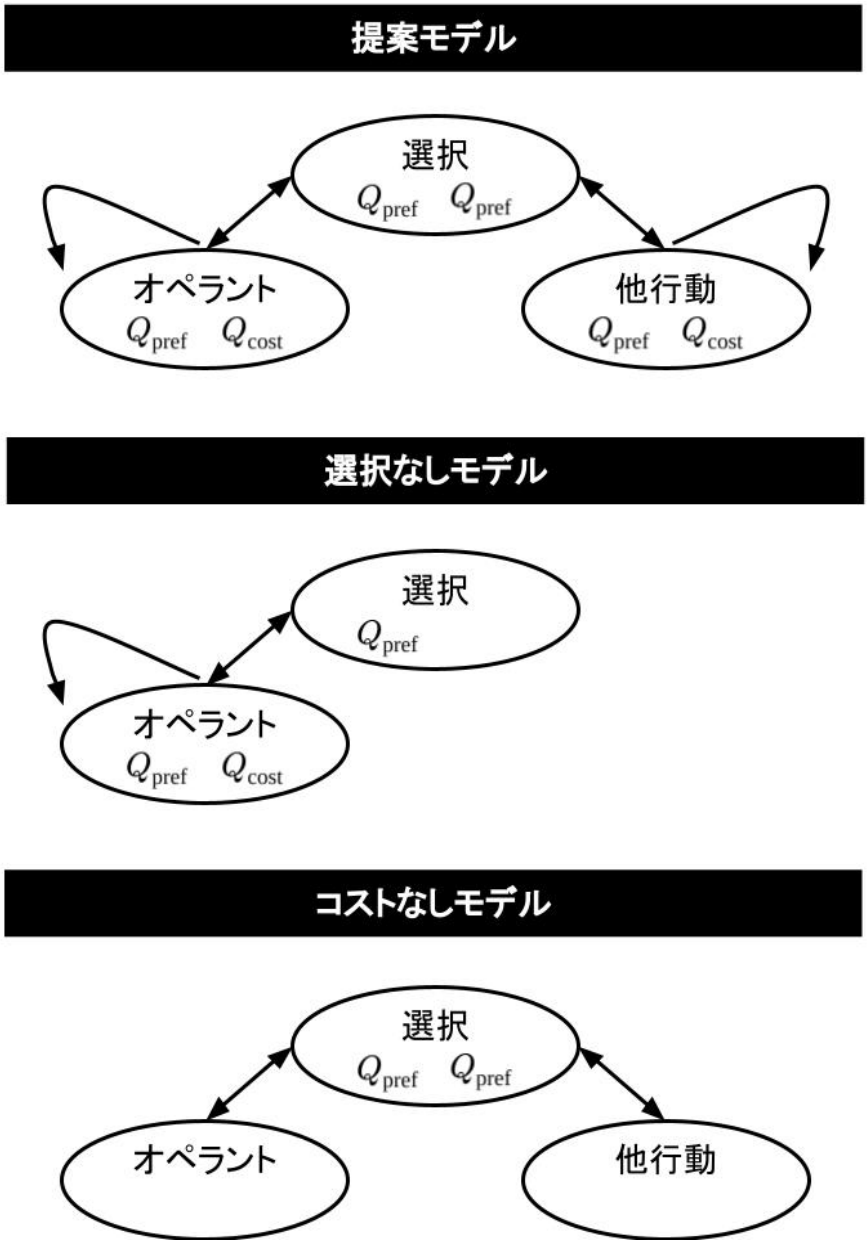

図 3.1.1. 提案モデルと 2 つのノックアウトモデルの概念図

楕円は状態を示し, 矢印は状態間の遷移を示す. 楕円内に記載されたパラメータに基づいて遷移確率を決定する.

提案モデルにおけるエージェントの振る舞いは以下の通りである. 「選択」において, エージェントはオペラント反応と他行動のいずれかを選択する. 選択の結果, 「選択」から「オペラント」または「他行動」のいずれかに遷移する. このとき, 各行動への選好 Q_{pref} に基づいて選択する. Q_{pref} の計算方法は次節で説明する. 「オペラント」では, エージェントはオペラント反応に従事し, 反応を自発するごとに「オペラント」にとどまるか, 「選択」に戻るかを, 「選択」への遷移コストを表す Q_{cost} に基づいて決定する. Q_{cost} の数学的定義は後述する. 「他行動」はエージェントがオペラント反応ではなく他行動に従事することを除けば, 「オペラント」と同じである.

エージェントは自身の行動と環境から与えられる結果に基づいて選好を表す Q_{pref} と切り替えに要するコストを表す Q_{cost} の双方を学習する. Q_{pref} はある反応によって得られる強化子の価値によって定義され, Q_{cost} はその反応に従事し始めてから強化子を得るまでに要する反応数によって定義される. Q_{pref} は 1 回のバウトごとの強化子の有無によって式 3.1 に従って更新される.

$$Q^i_{pref} \leftarrow Q^i_{pref} + \begin{cases} \alpha^+ \cdot \left(r^i_t - Q^i_{pref}\right) & \text{if reward is presented} \\ \alpha^- \cdot \left(0 - Q^i_{pref}\right) & \text{otherwise} \end{cases} \tag{3.1}$$

ここで α^+ と α^- は学習率であり, 報酬の有無によって異なる学習率を仮定している. Q^i_{pref} は「オペラント」と「他行動」のいずれかへの選好を表し, r^i_t はある時点において従事した反応 i に対して与えられた報酬の価値を表す. 「選択」では以上の Q^i_{pref} と以下の式に従って「オペラント」と「他行動」を選択する.

$$p_i = \frac{\beta e^{Q^i_{pref}}}{\sum_{i \in \{operan,\ other\}} \beta e^{Q^i_{pref}}} \tag{3.2}$$

ここで β は逆温度であり選択のランダムさを調整するパラメータである. β が高いほど選好が高い選択肢への選択確率が上昇し, 小さくなるにつれて選択が無差別になる.

Q_{cost} はある反応に従事し始めてから実際に報酬を得られるまでに要するコストであり, 「オペラント」あるいは「他行動」に遷移してから, 報酬が与えられるまでに生じた反応数に応じて, 式 3.3 に従って更新される.

$$Q^i_{cost} \leftarrow Q^i_{cost} + \alpha^+ \cdot \left(\ln x^i - Q^i_{cost}\right) \quad 3.3$$

ここで x^i は直前の遷移から報酬が与えられるまでに自発した反応数を表す. 他のパラメータは式 3.2 と同様である. 「オペラント」と「他行動」では式 3.1, 3.3 で定義された Q_{pref}, Q_{cost} に応じて, 一回の反応ごとに「選択」へと遷移するか, 現在の状態に留まるかを以下の式によって決定する.

$$p^i_{stay} = exp \frac{1}{w_{pref} Q^i_{pref} + w_{cost} Q^i_{cost}} \quad 3.4$$

ここで w_{pref} と w_{cost} は Q_{pref} と Q_{cost} への重み付けパラメータであり, どちらのパラメータも大きくなるにつれて, 現在の状態に留まる確率が上昇する. さらに, バウトの長さに対する影響は, 強化率のような選好に影響を与える要因より, コストに影響を与えるスケジュールタイプの要因の方が大きいため, ここで $w_{pref} \leq w_{cost}$ と仮定した.

シミュレーション

シミュレーションでは, バウト・休止パターンの研究で頻繁に採用される, VI スケジュールを採用した. VI スケジュールでは報酬が動物の反応とは無関係に時間経過によって準備され, 報酬が準備されてから初発の反応に対して報酬が与えられる. さらに報酬が呈示されてから次の報酬が準備されるまでの間隔は, 任意の平均間隔となるようにランダムに決定される. 具体的には任意の平均間隔を T としたとき, 期待値が T となるような指数分布から各間隔は生成される. さらに, VI スケジュールでは分布への収束性を担保するために Fleshler and Hoffman (1962) によって提案された方法を用いて, 疑似乱数としてサンプリングされる. 本シミュレーションも Flesher and Hoffman (1962) に基づいて VI スケジュールの間隔を決定した.

シミュレーションの流れは, まずエージェントはシミュレーション開始時には「選択」に滞在しており, Q^i_{pref} と Q^i_{cost} の初期値は 0 とした. その後, 次のタイムステップにおいてエージェントは式 3.2 に従って, 「オペラント」と「他行動」を選択し, その選択された状態へと遷移した. その遷移した先の状態において各タイムステップにおいて確率 $p = 0.33$ で, その状態に対応した反応を自発した. 一度, 反応を自発するたびにスケジュールに応じて強化子の有無が決定され, エージェントが現在の状態に留まるかが式 2.4 に従って決定された. 強化子が与えられた場合は, 式 3.1, 3.3 に従って Q^i_{pref} と Q^i_{cost} が更新されて, 「選択」へと遷移した. 強化子が与えられなかった場合は, 式 3.4 に従ってバウトの終了を決定し, 終了する場合には式 3.1 に従って Q^i_{pref} のみが更新されて「選択」へと遷移し, バウトが継続する場合に何もせずに現在の状態に留まるものとした. シミュレーションは VI 120 秒で行った.タイムステップを 0.1 秒として, エージェントが 1000 回の強化子を獲得するまで継続した.

シミュレーションに用いたパラメータは $\alpha^+ = 0.05$, $\alpha^- = 0.01$, $\beta = 12.5$, $w_{pref} = 1.0$, $w_{cost} = 3.5$ であった. さらにオペラント反応によって得られる強化子と他行動によって得られる強化子の価値はそれぞれ 1.0 と 0.5 とした. シミュレーションの実装と実行は全て Julia 1.0 によって行った. シミュレーションの実行には 1.80GHz Intel i7-8565 プロセッサ, 16GB RAM, 1TB SSD を搭載したコンピュータ (Ubuntu 18.04 LTS で動作)を用いた.

結果

図 3.1.2.の左図は, 提案モデルと 2 つのノックアウトモデルによって生成された反応列をラスタープロットである. 選択なしモデルでは, 反応が密で一様に分布しており, 休止期間が生じていなかった. コストなしモデルでは, 反応が疎で一様に分布しており, 反応が密に自発される期間が生じていなかった. 提案モデルでは, 反応が密に分布している期間が, 比較的長い休止期間によって隔てられていた. 図 3.1.2.の右図は, それぞれのモデルで生成された IRT の対数生存関数を示す. 選択

なしモデルでは, 急峻な一本の直線, コストなしモデルでは, 緩やかな直線でそれぞれ表されている. 提案モデルでは, 急峻な直線と緩やかな直線から構成される折れ曲がりのある曲線となった.

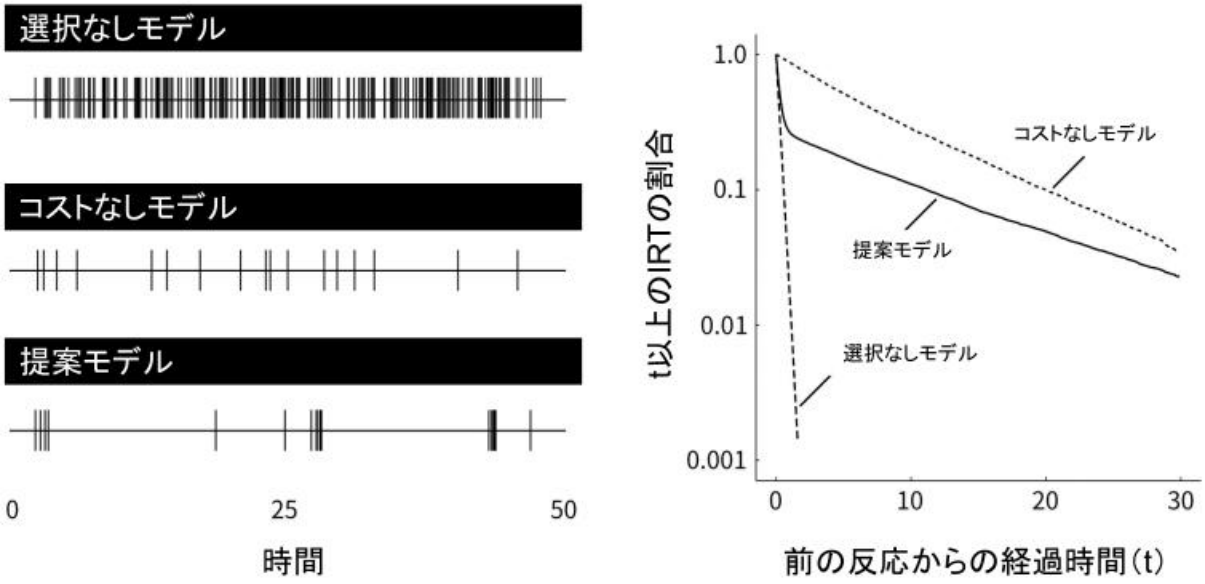

図 3.1.2. 各モデルによって生成された IRT のラスターと対数生存関数

2 つのノックアウトモデル (左上・中央) と提案モデル (左下) によって生成された反応をラスター表示した. 生存関数は任意の事象が特定の時間を超えて生じない確率を示すものであり, 右図は IRT の生存関数を描き, 縦軸を常用対数によって表したものである. 事象の発生間隔が指数分布に従うとき対数生存関数は一本の直線で表される.

実験 1 考察

図 3.1.2.が示すように, 提案モデルのみで, バウト・休止パターンが再現され, 2 つのノックアウトモデルではバウト・休止パターンの再現に失敗した. 本実験では, 選択とコストという 2 つのプロセスによってバウト・休止パターンが生成されていると仮定した. 選択なしモデルでは, エージェントがオペラント反応のみに従事することで, 常に反応が高頻度で自発され, バウト・休止パターンにおける休止が生じなかった. コストなしモデルでは, 選択に要するコストが存在しないため「オペラント」と「他行動」を高速で切り替えたことで, 反応が密に自発されるバウトが生じなかった. これらの結果から, 選択とコストという 2 つのプロセスは, それぞれが, 休止とバウトの生成を担っており, これらが組み合わさることで始めてバウト・休止パターンが再現される.

本実験はコンピュータシミュレーションであることから, 環境とエージェントは完全な制御下にあるため, 現実の動物を使用した研究とは異なり, 他の要因が交絡する可能性を完全に排除している. Smith et al. (2014) では, 選択とコストがバウト・休止パターンの背後にある可能性を示唆したが, 実験的に統制されていない他の要因や, あるいは生物が内在的に有する他の何らかのプロセスが関与する可能性があった. 従って, Smith et al. (2014) では選択とコストがバウト・休止パターンの生成に必要であることを示したが, それが十分であることまでは示していない. それに対して本実験で採用した構成論的なアプローチでは, これら 2 つのプロセスで十分にバウト・休止パターンが再現できることを示した.

バウト・休止パターンを生成するために重要なのは, モデルの特定のアーキテクチャではなく, 選択とコストのメカニズムである. 提案モデルは 3 つの状態と 5 つの方程式から構成されており, それらの方程式は Q-learning と呼ばれる最も有名な強化アルゴリズムの 1 つである. このようなモデルのアーキテクチャやアルゴリズムを他のものに置き換えたとしても, 選択とコストが実装されていれば, 他のモデルでもバウト・休止パターンを再現することができる. また, 式 3.2 のソフ

トマックス関数や式 3.3 中の対数など, 特定の式形式は他の形式への置き換えが可能である. 従って, この 2 つのプロセスを実装している限り, 他の形式の可能性を否定するものではない.

実験 2：提案モデルによる過去の実験結果の再現

実験 1 では, 提案モデルによってバウト・休止パターンが再現できることを示した. 実験 2 では, 様々な環境の下での提案モデルの振る舞いを分析する. 先行研究では強化率 (Shull et al., 2001, 2004), 遮断化レベル (Shull et al., 2001; Shull, 2004), 消去 (Brackney et al., 2011, 2012; Shull, et al., 2002), そしてスケジュールタイプ (Shull et al., 2001, 2004) の要因が検討されている. これらの実験で採用された実験環境をシミュレーション上で再現することで, 過去の研究と同様の結果が提案モデルから得られるか検討する.

モデル

実験 1 の提案モデルのみを使用した.

シミュレーション

提案モデルを用いて, 4 つの環境変数の内の 1 つを操作して, それ以外の変数については実験 1 と同じに設定した. スケジュールなど一部の条件を除いてシミュレーション手順は実験 1 と同様であった.

1) 強化率, 2) 遮断化レベル, 3) 消去, 4) スケジュールタイプの 4 つの実験操作をそれぞれ独立に適用した. 1) 強化率は, VI スケジュールの平均間隔を変化させることで操作した. このシミュレーションで用いた平均間隔は, VI 30 秒, 120 秒, 480 秒 (1 分間に 2.0, 0.5, 0.125 個の強化子) である. 2) エージェントの遮断化レベルを制御するために, 「オペラント」で得られる報酬の価値を変化させた. 報酬価値は 0.5, 1.0, 1.5 とし, それぞれ低遮断化, ベースライン, 高遮断化レベルに対応する. 3) VI 120 秒で 1000 強化の後に, 消去へとスケジュールを切り替えた. 消去は 3600 秒 (36,000 タイムステップ) 経過した時点で終了した. 4) スケジュール

タイプの操作では VI スケジュールに微小な VR スケジュールを連結スケジュールによって追加した. VI スケジュールの平均間隔は 120 秒に固定し, VR の値は 0, 4, 8 とした. 使用したモデルのパラメータ, 及びシミュレーションの実行環境は実験 1 と同様であった.

解析

消去の IRT データの解析のために, 二重指数分布の各パラメータが時間依存的に減少する動的二重指数分布 (Brackney et al., 2011) を使用した. 動的二重指数分布は式 3.5 によって表される.

$$p(IRT = \tau) = (1 - q_t)\omega_t e^{-\omega_t \tau} + q_t b_t e^{-b_t \tau} \quad 3.5$$

消去によって q_t, b_t, そして ω_t は時間とともに式 3.6, 3.7 に従って指数関数的に減衰する.

$$1 - q_t = (1 - q_0)e^{-\gamma t} \quad 3.6$$

$$b_t = b_0 e^{-\delta t} \quad 3.7$$

ここで, パラメータ γ と δ は, それぞれ q_t と b_t の減衰率を表す. モデルパラメータ q_t, b_t, ω_t のいずれの減衰によっても反応が完全に消去される可能性があるため, どのパラメータが実際に減衰したかを確認する必要がある. ω はシミュレーション中に 1/3 に固定されたため, ここでの解析からは除外した. そこで, q_t と b_t のどちらか一方, あるいは両方が減衰したかどうかを確認するために, *qb*-減衰モデル, *q*-減衰モデル, *b*-減衰モデルの 3 つのモデルを WAIC (widely applicable information criterion; Watanabe, 2010) によって比較した. 事後分布の推定には Stan (Carpenter et al., 2017) による Markov chain Monte Carlo (MCMC) を用い, そこで生成された MCMC サンプルを使用して WAIC を計算した.

結果

図 3.2.1.に, 4 つのシミュレーションによって生成された IRT の対数生存関数を示した. 強化率の減少によって右側の直線の傾きが緩やかになり, 切片が上昇した (図 3.2.1 左上). 遮断化レベルの減少によって右側の直線の傾きが緩やかになり,

切片が上昇した (図 3.2.1 右上). 図 3.2.1 の左下のパネルは消去中のデータを 20 分おきに三分割したものとベースラインでの IRT のそれぞれで対数生存関数を描いている. ベースラインから消去の進行に伴い右側の直線の傾きが緩やかになり切片が上昇した. VI スケジュールに微小な VR スケジュールを追加することで, 追加した VR の大きさに応じて切片が減少したが傾きに変化はなかった (図 3.2.1. 右下).

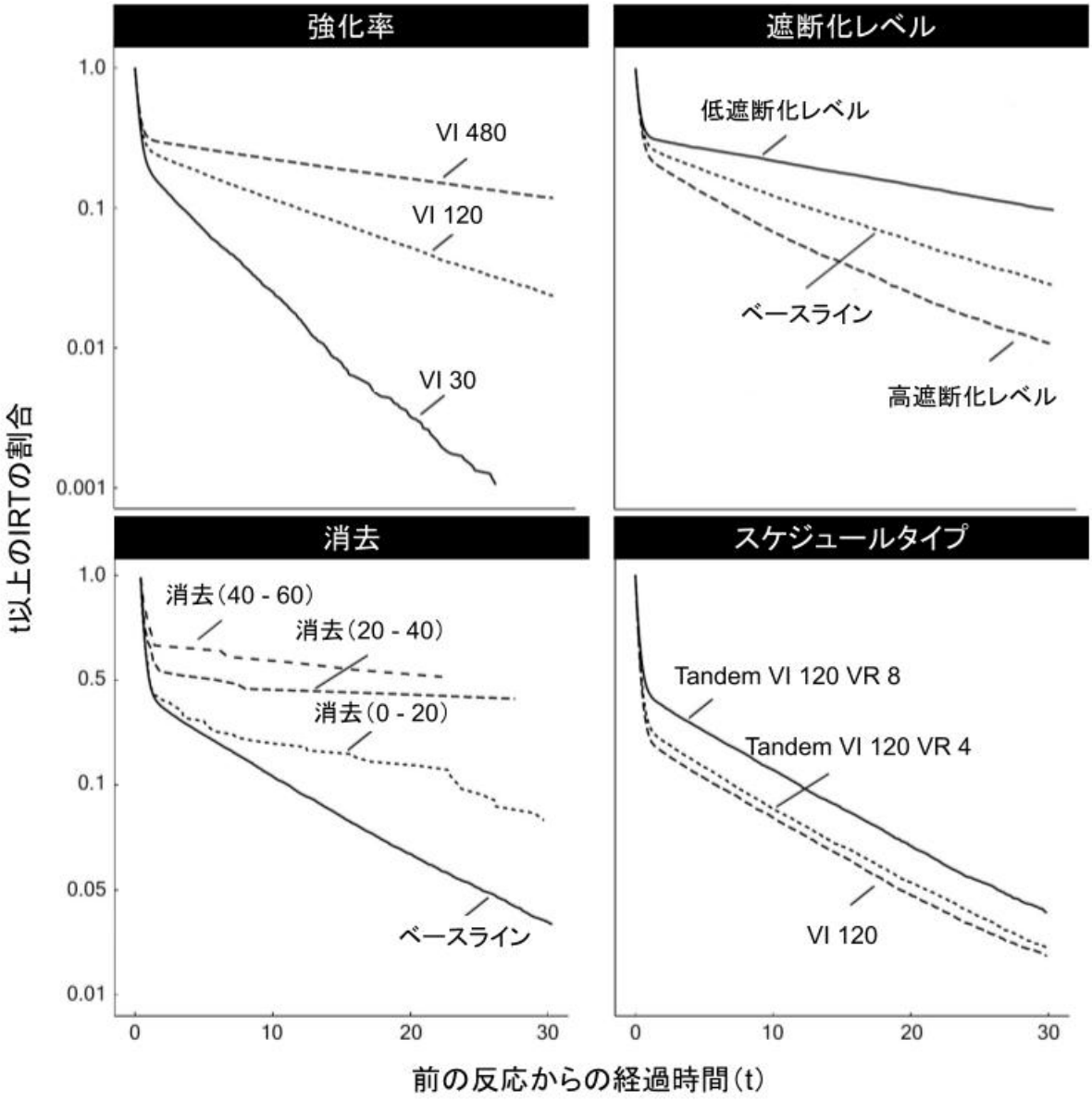

図 3.2.1. 強化率, 遮断化レベル, 消去, そしてスケジュールタイプがバウト・休止パターンに与える影響

提案モデルの挙動を異なる実験操作の下でシミュレートして生成された IRT によって対数生存関数を描いた.

消去を除いた全てのシミュレーションの IRT データに二重指数分布を当てはめてパラメータ推定を行った. パラメータは $\omega, b,$ そして q の 3 つであり, ω はバウト内反応率, b はバウト発生率, そして $\frac{1}{1-q}$ はバウト長を表す. 推定されたパラメータをシミュレーションごとに表 3.2.1.にまとめた. 強化率を操作した場合には, その減少に伴って ω と q が上昇し, b は減少したため, バウト内反応率が上昇した一方でバウト発生率とバウト長は短くなった. 遮断化レベルの操作は強化率の操作と同様の結果であり, バウト内反応率は上昇し, バウト発生率とバウト長は減少した. スケジュールタイプの影響は, バウト内反応率では条件によって異なるものの一貫した傾向は認められず, バウト発生率は変化せず, バウト長は追加の VR の長さに応じて上昇した. 消去のデータには動的二重指数分布を当てはめて, 二重指数分布のうち q と b のいずれか, もしくは双方が減少する 3 つのモデルを WAIC によって選択した. モデルごとの WAIC を表 3.2.2.にまとめた. WAIC の比較からq と b の双方が減衰するモデルが選択されたことから, 消去に伴ってバウト発生率とバウト長の双方が減少した.

表 3.2.1. 各実験操作の条件ごとの二重指数分布のパラメータ

実験操作	条件	ω	b	q
強化率	VI 30	3.08	0.23	0.17
	VI 120	3.06	0.09	0.27
	VI 480	3.19	0.03	0.31
遮断化レベル	低	3.04	0.24	0.17
	中	3.15	0.08	0.26
	高	3.17	0.03	0.31
スケジュールタイプ	VI 120	3.07	0.09	0.26
	Tandem VI 120 VR 4	3.23	0.08	0.19
	Tandem VI 120 VR 8	3.11	0.08	0.16

表 3.2.2 動的二重指数分布の WAIC

Model	WAIC
qb-減衰	1.936
b-減衰	1.94
q-減衰	1.98

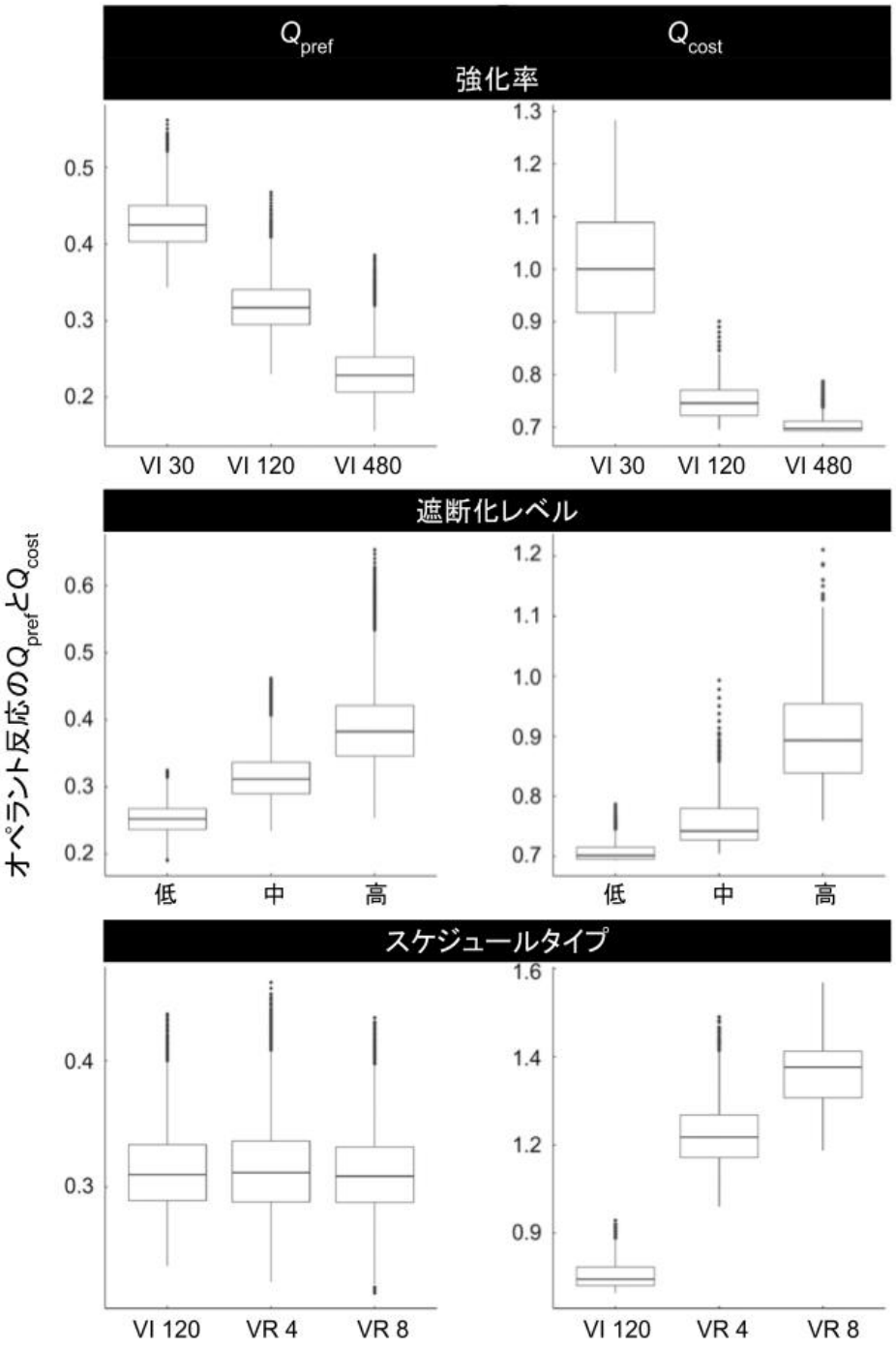

図 3.2.2. 強化率, 遮断化レベル, スケジュールタイプの条件ごとのQ_{pref}とQ_{cost}

学習が収束し, 安定したデータを使用するために, 500 強化から 1000 強化を解析区間とした.

図 3.2.2.は消去を除く 3 つのシミュレーションでの Q_{pref} と Q_{cost} の値を箱ひげ図で示している. これは最終的な IRT という行動指標として顕れる環境操作が, モデルの内部状態の内どのプロセスに媒介されているものかを表す. 強化率を操作したシミュレーションでは強化率の減少に伴って Q_{pref} と Q_{cost} が減少している. 遮断化レベルの影響は遮断化レベルが増加するに従って, Q_{pref} と Q_{cost} が増加していた. スケジュールタイプの操作は Q_{pref} には影響が一切なかったが, Q_{cost} は追加の VR の大きさに依存して上昇した.

実験 2 考察

実験 2 では, 提案モデルが, 過去の研究で報告されていた動物の行動と同様の振る舞いをするか検証した. 強化率, 遮断化レベル, 消去, そしてスケジュールタイプの 4 つの実験操作の影響を検討した. 強化率, 遮断化レベル, そして消去はバウト開始率とバウト長に影響を与え, スケジュールタイプはバウト長にのみ影響を与えた (図 3.2.1.; 表 3.2.1.; 表 3.2.2.). これらの研究は, 過去の動物実験で得られた結果と概ね一致する (Brackney et al., 2011, 2012; Shull et al., 2001, Shul et al., 2002; Shull et al., 2004; Shull, 2004). 過去の動物の実験でも, 対数生存関数によって行動の変化を評価していた. 多くの実験で, IRT は対数生存関数上の折れ曲がりのある 2 本の直線によって示されており, 環境操作による対数生存関数の変化は, 本研究の結果と一致するものであった. 提案モデルはバウト・休止パターンの再現に加えて, その振る舞いが動物実験の知見と一致することは, その生成機構のモデルとしての妥当性を示すものである.

提案モデルは Q_{pref} と Q_{cost} という内部状態を仮定しており, 図 3.2.2.に示されるように実験操作に対して異なる振る舞いを示す. Shull (2001) では強化率, 遮断化レベル, そして消去を動機付け操作としてカテゴライズしているが, これらの操作によるバウト・休止パターンへの影響の一貫性や, 図 3.2.2.に示されるように, 内部状態への影響が類似しているため, これら異なる実験操作が, 行動出力へ与える影響は同一であると考えられる. さらに図 3.2.2.から Q_{pref} と Q_{cost} の関係性は必ず

しも独立していないことが推測できる. 強化率と遮断化レベルの操作では Q_{pref} の上昇によって Q_{cost} も上昇しているが, スケジュールタイプへの操作の結果から分かるように Q_{cost} の上昇は Q_{pref} に影響を与えない. 従ってこれらの内部状態は Q_{pref} から Q_{cost} への一方向性の影響を持つ. この理由は式 3.4 で現在の状態に留まる確率を Q_{pref} と Q_{cost} によって算出しているためである. Q_{pref} が上昇すると「オペラント」への滞在確率, つまりバウト長が上昇する. Q_{cost} はバウト開始から報酬が与えられるまでに自発した反応数に依存するため, Q_{pref} に由来するバウト長の上昇は結果的に Q_{cost} の上昇を招く. 従って強化率や遮断化レベルの操作による Q_{cost} への影響は Q_{pref} を媒介した二次的な作用である.この二次的な Q_{cost} への影響は実験的に操作可能であり, 低反応率分化強化スケジュール (differential reinforcement of low response rate; DRL) によって, バウト内反応のような短い IRT を強化しないスケジュールで実現可能である. 例えば tandem VI DRL スケジュールのように設計することで, VI 成分で強化率を操作し, DRL 成分でバウト内反応への強化を防ぐことで, バウト・休止パターンに対して強化率が与える純粋な影響を検証できるだろう.

研究 3 総合考察

研究 3 では, 強化学習によってバウト・休止パターンが生成される過程をモデル化し, コンピュータシミュレーションと実験結果を比較することにより, その妥当性を検討した. モデルでは, バウト・休止パターンを生成するために, オペラント反応と他行動の選択と, 反応の切り替えのコストという 2 つの独立したメカニズムを仮定した. 実験 1 では, 提案モデルが VI スケジュール下でバウト・休止パターンを再現できることが示された. 実験 2 では, 様々な実験操作と提案モデルにおけるバウト・休止パターンの構成要素との間に一貫性があることを確認した. これらの結果は, 選択とコストによって反応がバウト・休止パターンとして表れるという仮説を支持するものである.

Kulubekova and McDowell (2008) は Skinner (1981) の「結果による選択」を遺伝的アルゴリズムによって実装した計算論的モデル (2004) によってバウト・休止パターンの再現を試みた. Kulubekova and MacDowell (2008) によればバウト・休止パターンの再現に成功したが, 実際には IRT の分布は対数生存関数上で, 2 本の直線から構成される曲線ではなく, 明確な折れ目のない緩やかな曲線であった. これはシミュレーションで得られた IRT 分布は実際の動物のデータとは異なり, バウト・休止パターンを必ずしも再現できていないことを意味する.

大脳基底核の神経回路は, ハイパー直接路, 直接路, そして間接路という 3 つの主要な経路によって構成されている. 直接路と間接路は, 線条体から異なる経路を介して大脳基底核の出力を担う淡蒼球内節 (internal segment of globus pallidus; GPi), 黒質網様部 (substantia nigra pars reticulata: SNr) へと達する. 直接路は, 線条体のドーパミンの D1 受容体をもつニューロンが, GABA 作動性の抑制性の投射を GPi, SNr へと行う経路である. 間接路は, ドーパミン D2 受容体を持つ線条体のニューロンが淡蒼球外節 (external segment of globus pallidus; GPe) を介して, GPi, SNr へと間接的に投射する経路である. 間接路の投射は全て GABA による抑制性の投射である. 従って線条体から GPe への抑制性投射は結果的に GPi, SNr が興奮する. GPi, SNr は抑制性であるため, 直接路は出力先の視床や上丘を興奮させ, 間接路は抑制する. これらの拮抗的な支配によって運動の開始や終端を制御している (Nambu, 2004).

Nonomura et al. (2018) では, 確率的な選択課題において直接路と間接路が報酬と無報酬という異なる結果を符号化していること, そして光遺伝子によるそれぞれの経路選択的な操作によって, 次の試行で前の選択から変えるか, そのまま現在の選択に滞在するかを, 独立に制御していることを明らかにした. Nonomura et al. (2018) は離散試行型の選択課題であり, フリーオペラント事態とは大きく異なる実験系であるため, バウト・休止パターンが同様のメカニズムによって担われているかは定かではない. しかし, オペラント反応同様にバウト・休止パターンを示

すリッキングでは, SNr から上丘の経路選択的な興奮によってリッキングを止めることが報告されていることから (Rossi et al., 2016; Toda et al., 2017), オペラント反応においても同様の回路が関与していることが考えられる. 提案モデルでは選択とコストという 2 つのプロセスによってバウト・休止パターンが生成されているとした. モデル上のそれぞれのプロセスは Nonomura et al. (2018) の意思決定課題における直接路, 間接路の機能, つまり反応の結果による選択と滞在とスイッチの意思決定, と類似していることから, オペラント反応のバウト・休止パターンにおいても, 同様の回路が重要な役割を果たす可能性が示唆される.

大脳基底核の回路以外にも, 研究 3 で想定したメカニズムに対応する神経基盤が存在し, 前帯状皮質背側部は, 現在従事しているタスクを継続するか, あるいは他のタスクへと移行するか, という意思決定を担っていると考えられている (Shenhav et al., 2016). そこでは, それぞれのタスクから得られる報酬の期待値と, 遷移に要するコストを考慮して, タスクの継続と移行の意思決定を行っているとされており, 提案モデルで想定した, 選択とコスト, という 2 つのメカニズムの双方を反映したものである (Shenhav et al., 2016). このことから, 前帯状皮質背側部がバウト休止・パターンの生成に関与している可能性が考えられる.

研究 3 では, バウト・休止パターンと呼ばれるオペラント反応の時間構造が生成される過程を強化学習によってモデル化した. 提案モデルの最大の特徴は, 動物の行動全体を, オペラント反応と他行動との選択行動として捉えることにある. 従来の行動分析学と学習心理学では, わずかな研究でのみ, 実験的に定義された行動以外の行動に注目していた (Staddon and Simmelhag, 1971). 本研究では, その他行動を明示的にモデルと組み込み, 動物の行動をオペラントと他行動という 2 つの状態間の遷移として記述することで, オペラント反応が示すバウト・休止パターンと呼ばれる, 時間的な構造が説明できることを示した. さらに, 提案モデルでも選択とコストという 2 つのプロセスを仮定して, 機械論的なモデルを構築することによって, バウト・休止パターンを生み出す神経科学的な作業仮説として扱

うことができる. このように研究 3 では, 他行動を明示的に取り入れたモデル化によって, 既存の研究では注目されることが少なかった他行動の重要性を強調すると同時に, 神経科学との接点を作り出すことに成功した.

研究 4：行動ネットワークの構造変化としての習慣形成

背景と目的

研究 3 では, 他行動を状態として導入し, 動物の行動をオペラントと他行動間の遷移とみなすことで, バウト・休止パターンの再現に成功した. しかし, 現実の生物の行動は, オペラント反応と他行動という, 実験者にとって都合よい単純な二分法によって区別できるようなものではなく, 現実には無数の反応が存在するのみである. この事実に着目して, 研究 3 で単一の状態として扱った他行動を, 複数の反応へと分解して, あらゆる反応が相互結合したネットワークとして行動を捉え直す. ここでは, 実験事態において観測される, あらゆる反応をネットワークへと組み込むことを想定しており, その時間スケールは, 数十分から数時間に渡る. こうした長い時間スケールで行動を観測することによって得られる, 無数の反応とその間の遷移は, ネットワークとして表現可能であり, 行動科学にネットワークの構造とそれによってもたらされる行動の特性という新たな視点を導入する. 研究 4 では, 単一の反応が持つ, 目的志向性や習慣といった特性が, 個々の反応に内在する, あるいは反応に対する直接的な制御系ではなく, ネットワークという行動のマクロな構造によって説明できることを, シミュレーションによって示す.

所与の環境の下で柔軟に振る舞うために, 動物は自身の行動の結果に基づいて行動を選択する必要がある. このような行動を目的志向行動 (goal-directed behavior) という. ある状況下で, 同じ行動を繰り返すことで, その行動が結果ではなく状況によって引き出されるようになる. このような行動を習慣 (habit) と呼ぶ. 目的志向行動は, 生物が自分の外部環境に関する情報に基づいて, 自分の行動が環境にどのような影響を与えるかを判断する必要があるため, 多くの計算コストが要求される. これに対し, 習慣はよりパターン化され, 柔軟性に欠ける行動である

代わりに, 要する計算コストも少なくなる. この意味で, 習慣形成は, 動物による計算コストの最適化過程と見なすことができる.

習慣形成の初期の研究は, 学習心理学で盛んに行われてきた (Adams and Dickinson, 1981; Dickinson, 1985; Dickinson et al., 1983). そこでは, ある反応が目的志向行動か, 習慣かは, オペラント反応の報酬価値への感受性によって定義された. 報酬低価値化手続きによって感受性の有無が検証される. まずフリーオペラント事態で, オペラント反応を任意の報酬によって訓練を行う. 反応が獲得された後に, その反応の訓練で用いた報酬を味覚嫌悪学習 (Adams and Dickinson, 1981; Garcia et al., 1966) や, 実験前給餌による飽和化 (Balleine and Dickinson, 1998) によって報酬の価値を低減させる. その後, フリーオペラント事態へと戻り, 動物が低価値化された報酬を得るために反応をするか検証する. ここで動物が反応しなければその反応は目的志向行動とされ, 反応を続ければ習慣とされる (Dickinson, 1985). ここで習慣形成とは反応とその結果による行動の制御, つまり反応と報酬 (R-O 連合) による制御から, 刺激と反応 (S-R 連合) による制御への移行とみなされている (Dickinson, 1985). この S-R から R-O への行動制御の移行は, 強化学習におけるモデルフリー学習とモデルベース学習として捉え直すことができる (Daw et al., 2005). ヒトを対象とした実験では, 2 段階マルコフ決定課題の選択パターンを, このモデルフリーとモデルベース学習の競合として捉えて, 習慣形成の研究がなされている (Daw et al., 2011). このように, 基準的な見解では, 2 つの異なる行動の制御系が, 単一の反応に対して, 競合的に制御を及ぼしているとされている (Daw et al., 2005, 2011; Perez and Dickinson, 2020).

しかし, 一部の研究者は, 習慣形成に関する既存の研究を再検討し, 習慣となった反応でも, 結果によっても制御される可能性を示すことで, 習慣形成の規準的な理論に疑問を投げかけている (De Houwer, 2019; Kruglanski and Szumowska, 2020). Dezfouli and Balleine (2012) は, 規準的な見解とは対照的に, 2 つのシステムの競合的な制御ではなく, それらは階層的に組織されていることを提案してい

る. そうした理論の下で, 習慣形成を反応列の形成と見なす, という新たな提案をした. 彼らのモデルでは, エージェントは目的志向的に目標を選択し, そこに到達するための反応列を生成する. Dezfouli と Balleine (2012, 2013, 2014) によって報告された一連の研究では, 二段階マルコフ決定課題を主に扱い, 報酬感受性に関する実験のごくわずかしか扱っていない. さらに Garr et al. (2019) は, 二段階マルコフ決定課題で報酬低価値化手続きを行っており, そこでは反応列は獲得されたものの, 報酬への感受性は失われていないことを報告しており, 反応列の獲得を本来の意味での習慣として捉えることに疑問を投げかけている. 従って, 習慣形成に関する新しい見解の妥当性を確認するためには, まだ多くの理論的・実験的な証拠の積み重ねが必要であることを示唆している.

2つの行動の制御系の競合という基準的な見解は, フリーオペラント事態での習慣形成 (Daw et al., 2005; Perez and Dickinson, 2020) とヒトを対象とした 2 段階マルコフ決定課題 (Daw et al., 2011) の双方の実験系を扱っている. その一方で, 2つの制御系の階層的な制御と反応列の形成という習慣形成の見方は, ヒトを対象とした 2 段階マルコフ決定課題への適用が主であり, フリーオペラント事態への適用は限られている. 本研究では, 2 つの制御系の競合という見方ではなく, Dezfouli and Balleine (2012) で提案されたような, 階層的な制御系とその下での反応列形成という見方を, フリーオペラント事態へと適用する. そして, 報酬価値感受性の欠如という, 古典的な習慣が, 反応列の形成という見方でも説明できることをシミュレーションによって明らかにする.

行動ネットワーク

学習心理学では実験的に厳密に統制された環境下で, 実験的に定義された動物の反応 (レバー押し, キー突き, ノーズポーク, 恐怖刺激へのすくみ, 唾液分泌, 水を舐める, まばたきなど)を測定する. しかし動物は, 実験的に定義された反応だけに従事するのではなく, 様々な反応 (毛づくろい, 探索, 歩行など) に従事する (Skinner, 1948; Staddon and Simmelhag, 1971). 近年の機械学習の進歩により,

行動の詳細な構造を客観的に測定できるようになった (Markowitz, et al., 2018; Mathis, et al., 2018; Wiltschko, et al., 2020). こうした手法は主に神経科学で発展しており, 心理学, 神経科学, 行動生態学という異なる領域を統合して動物の行動を明らかにすることが期待されている (Datta, et al., 2019; Leon, et al., 2021). その一方で, 従来の神経科学や心理学における行動の見方や理論は, こうした新たな行動の定量化技術が発展する以前のアプローチから得られた知見に基づいている. そこで, 個々の反応が従う学習による変化という従来の視点に加えて. それらの行動が織りなす巨視的な構造という新たな視点を取り入れる. そして, その構造に由来する行動の性質として, 従来の行動的現象を説明する新たな理論的枠組みを提示する.

学習心理学の従来の研究では, 動物の実験的に定義された反応のみを測定し, 動物が実際に従事している他行動は無視されていた. しかし, 行動分析学では実験的に定義された反応に対して他行動が影響を与えることは古くから知られている. 例えば, 報酬提示直後に標的反応以外の特定の反応をするスケジュール誘導性行動 (Falk, 1966; Hymowitz, 1971; Gentry, 1968; Levitsky and Collier, 1968), 実験と無関係な反応に従事すること (Skinner, 1948), そうした行動が報酬提示間に特定のシーケンスで表れること (Staddon and Simmelhag, 1971), 学習された反応をそれ以外の反応が阻害する本能的逸脱 (Breland and Breland, 1961) などの例がある. さらに理論的には, オペラント反応のいくつかの特徴は, 他の反応の存在を仮定することで説明される (Guthrie, 1930; Herrnstein, 1970; Killeen and Fettermann, 1988; Yamada and Kanemura, 2020). これらの実験的事実と理論的要請は, 動物の反応が単独で存在するのではなく, 他の反応と関連して存在することを示している. 本研究では, こうした事実を踏まえて, 動物の行動を, 反応が相互結合したネットワークとしてモデル化する.

ネットワーク科学は, 1990 年代半ばから後半にかけて登場し, 幅広い分野に広がっている (Barabasi, 2016). ネットワーク科学の重要なポイントのひとつは,

個々のノードのつながりではなく, 巨視的なネットワークの構造を扱うことである. 例えば, 個々のノードがランダムに接続されたネットワークでは, 各ノード間の距離が大きく, 情報の伝達が遅くなるが, ネットワークの中にハブと呼ばれる, 他のノードから多くのエッジを獲得したノードがあると, そのノードを経由することで情報が高速に伝達されるようになる. 現実的な例としては, インフルエンサーが SNS で情報を発信することで, より多くのユーザーの目に留まり, 急速に情報が拡散していくようなものである. このように, ネットワークの構造はシステム全体の振る舞いと密接に関係している. ここでは, このようなネットワーク構造の視点を行動科学に応用する. 具体的には, 一つ一つの反応をノードとし, 行動を相互に接続されたノードのネットワークとして捉えることである. これにより, 既存の行動現象を, 行動の全体構造という新しい視点から説明しようとするものである.

ここでは, 行動ネットワークの計算論的定式化を行い, 習慣形成をネットワーク構造の変化という観点から説明する. 実験 1 では, 任意のネットワークを生成し, どのような構造で習慣が形成されるかを検討することで, 特定の反応にエッジが集中することで習慣形成が生じることを示した. 実験 2 では, 既存の行動研究で明らかになった, 習慣形成の促進・抑制要因が, 提案モデルでも同様の効果を持つかどうかを検討した. 習慣形成に関する重要な要因は以下の 3 つである. 1) 訓練量 (Adams, 1982; Dickinson, et al., 1995), 2) 報酬のスケジュール (Dickinson et al., 1983), 3) 選択の有無 (Colwill and Rescorla, 1985; Kosaki and Dickinson, 2010) である. これらの要因が提案モデルに及ぼす影響は, 既存の実験結果と一致した. さらに, 実験 3 では習慣形成に関する心理学的な仮説である, 反応-報酬相関説と反応-報酬接近説が異なる予測をする実験事態を提案する. これらの実験を通して, 他行動という行動分析学的な視点を学習心理学で扱われてきた現象に取り入れることで, これまでと異なる方法で現象を説明できることに加えて, 学習心理学で提案された仮説を検証することが可能な新たな実験の提案という貢献もたらす.

実験 1：オペラント反応へのエッジの集中によって習慣形成が生じた

モデル

エージェントの行動を, レバー押しや毛づくろい, 探索など異なる反応と各反応間の遷移から構成されるネットワークとみなす. 個々の反応がノードであり, その間での遷移がエッジとなる. 従ってエージェントの行動は, 総体としてネットワーク上での, あるノードから別のノードへの遷移の集合体として捉えられる. このネットワークの構造はエージェントが環境との相互作用を通して学習するものとして, ある反応から別の反応へと遷移した際に獲得できる報酬の価値に基づいて決定される. ここで用いられる学習則とネットワーク生成則, シミュレーションにおけるエージェントが従う行動則は以降のパラグラフにて詳細に解説する.

ネットワークの生成は任意の遷移間が持つ行動価値関数に基づいて行う. 行動価値関数は全ての反応の組み合わせごとに存在するため, エージェントが従事しうる反応のカテゴリー数を N としたとき, $N \times N$ の行列となる. 行動価値関数は, ある時点 $\mathrm{t}-1$ に従事していた反応を a_{t-1} として次の時点 t で a_t に遷移したときに得られる報酬 $r(a_{t-1}, a_t)$ に応じて, 式 4.1 に従って更新される.

$$Q(a_{t-1},\, a_t) \leftarrow Q(a_{t-1},\, a_t) + \alpha \cdot \big(r(a_{t-1}, a_t) - Q(a_{t-1},\, a_t)\big) \qquad 4.1$$

式 4.1 によって計算される行動価値関数に基づいて, 任意 2 つの反応 i, j 間にエッジが貼られる確率決定される.

任意の 2 つのノード間にエッジが付く確率は行動価値関数に依存し, 任意の反応 i からそれ以外の反応への遷移が持つ行動価値関数の比例配分によって表されるため, 任意の 2 つの反応 i, j 間にエッジが貼られる確率は以下のように表される.

$$p_{i,j} = \frac{Q(i,j)}{\sum_{i=1}^{N} Q(i,j)} \qquad 4.2$$

式 4.2 に従って全てのノードが少なくとも 2 つのエッジを持つようにエッジをサンプリングした. ネットワークの生成には Python のネットワーク解析ライブラリである networkx (Hagberg et al., 2008)を用いた.

次にシミュレーションにおけるエージェントの行動則について解説する. シミュレーションには大きく分けて 3 つのフェイズが存在する (各フェイズの詳細は次のセクションにて解説する). 1 つ目は訓練フェイズであり, ここでエージェントは, ある実験条件を模した環境との相互作用を通してネットワークの構造を学習する. 2 つ目はベースラインフェイズであり, ここでは獲得されたネットワークに従って様々な反応に従事する. 最後はテストフェイズであり, これは報酬低価値化手続きを模したものであり, 報酬の価値を除いて基本的には直前のベースラインフェイズと同様である. これらのフェイズの内, 訓練フェイズとそれ以外の 2 つのフェイズでは行動則が異なる.

学習フェイズではエージェントが任意の環境の下でネットワークの構造を学習する. ここでの行動則はランダムな反応の選択である. 各時点においてエージェントは何か 1 つの反応に従事するが, エージェントが従事する反応は直前の反応に関わらず, 全ての反応の中から等確率でサンプリングされるものとした.

ベースラインフェイズとテストフェイズでは, 報酬の価値に基づいて全反応から 1 つの反応を選択する. ある反応 i が選択される確率は以下の式によって決定される.

$$p_i = \frac{r_i}{\sum_{i=1}^{N} r_i} \tag{4.3}$$

r_i は反応 i によって得られる報酬の価値を表している. 式 4.3 に従って反応を選択した後に, 現在従事している反応からその反応へと遷移する. ここで 2 つの反応間の遷移にはネットワーク上における 2 つの反応を結ぶ経路の内, 最短となる経路によって行われた. 最短経路探索アルゴリズムのダイクストラ法 (Dijkstra, 1959) によってその経路を求めた. ここで任意の 2 つのノード s,t 間の距離は, ノード s からノード t までに介するエッジの数と定義した. 従って任意の 2 つのノードが直接的に接続されている場合, 最短距離は 1 となり, 間接的に接続されている場合には仲介するノードが増えるに従って増加する. エージェントはダイクストラ法によって求められた最短経路に含まれる反応を順に従事することで, 式 4.3 で選

択された反応へと到達する. その後, 式 4.3 によって次の反応を選択することで, 以上の手順を繰り返す. 最短経路探索には networkx (Hagberg et al., 2008) を使用した.

シミュレーション

実験 1 では任意のネットワークを生成して, その下で習慣形成が生じるかを検討する. 従ってエージェントが直接的に環境と相互作用を通して行動価値関数を学習する代わりに, 任意の行動価値関数を与えて, その下で式 4.2 に従ってネットワークを生成する. その後, ベースラインフェイズとテストフェイズを行い, 習慣形成が生じているか検討する.

ここではオペラント反応への集中度合いを系統的に操作できるように, 行動価値関数を定義した. 行動価値関数は反応カテゴリー数を N とした時に $N \times N$ の行列となる. そこで行動価値関数を疑似的に生成する上で, 長さ N のベクトルを $\boldsymbol{q}$ を用いた. $\boldsymbol{q}$ の 0 番目の要素をオペラント反応として (以下 Q-operant), それ以外の要素を他行動とした. $\boldsymbol{q}$ の直積 ($\boldsymbol{q} \times \boldsymbol{q} = \boldsymbol{q}\boldsymbol{q}^T = Q$)を行動価値関数とした. そして Q-operant を系統的に操作することでオペラント反応への集中度合いを操作する. シミュレーションでは Q-operant を 0 - 1.0 の範囲で変化させ, 他の要素の値を 0.001 に固定して行動価値関数を疑似的に生成し, 式 4.2 に従ってネットワークを生成した.

ベースラインフェイズではエージェントは式 4.3 と最短経路探索によって任意のネットワーク上を遷移する. ここでは最初に従事する反応を, 全反応の中でランダムに選択する. 次に, 式 4.3 に従って反応を 1 つ選択する. 報酬価値 $\boldsymbol{r}$ は全反応によって得られる報酬の価値をベクトルで表したものであり, 0 番目の要素がオペラント反応によって得られる報酬の価値, それ以外の要素は他行動によって得られる報酬の価値を表す. ベースラインフェイズでは $\boldsymbol{r}_0 = 1$として残りの要素を 0.001 とした. 反応が選択された後に最短経路を辿って, 選択された反応まで遷移する. その反応まで遷移した後に, その反応を起点として式 4.3 によって反応を再

度選択し, そこまでの最短経路を求めた. ベースラインでは以上の流れを 500 回繰り返して, 全反応の内にオペラント反応が占める割合を算出した. テストフェイズではベースラインフェイズと同様の手続きでシミュレーションを行ったが, 報酬低価値化手続きを再現するために, $\boldsymbol{r}_0 = 0$とした. 従って, テストフェイズにおいては式 4.3 によってオペラント反応が選択されることはない.

シミュレーションの実装と実行は全て Python 3.8.10 によって行った. シミュレーションの実行には 1.80GHz Intel i7-8565 プロセッサ, 16GB RAM, 1TB SSD を搭載したコンピュータ (Arch Linux 上で動作) を用いた.

結果

Q-operant を系統的に操作することで, オペラント反応へのエッジの集中度合を操作し, それが習慣形成に与える影響を検討した. Q-operant を 0.0 – 1.0 の範囲で変化させたところ, Q-operant が増加するに従って, ネットワーク上に存在するノードは徐々にオペラント反応へと接続されるようになり, Q-operant が 1.0 ではほぼ全てのノードがオペラント反応へと接続された (図 4.1.1. A, B の中央図). Q-operant の増加に伴い報酬低価値化への抵抗性は増加しており, 報酬が低価値化しているにもかかわらず, エージェントはオペラント反応に従事していることが示された (図 4.1.1. B の左図). オペラント反応にエッジが集中することで, オペラント反応の媒介中心性, 任意の 2 つのノードの最短経路にオペラントが含まれる確率が上昇している(図 4.1.1. の右図). これはエージェントがオペラント反応を選択するかしないかに関わらず, 任意の 2 つの反応間を遷移する経路にオペラントが含まれるやすくなることを意味している. さらにオペラント反応へとエッジが集中することで, ネットワークの平均距離, ネットワーク内の任意の 2 つのノードの最短経路の平均値, が縮小しており, その結果として反応間の遷移が効率化されて, シミュレーションに要する時間が短縮された(図 4.1.2.).

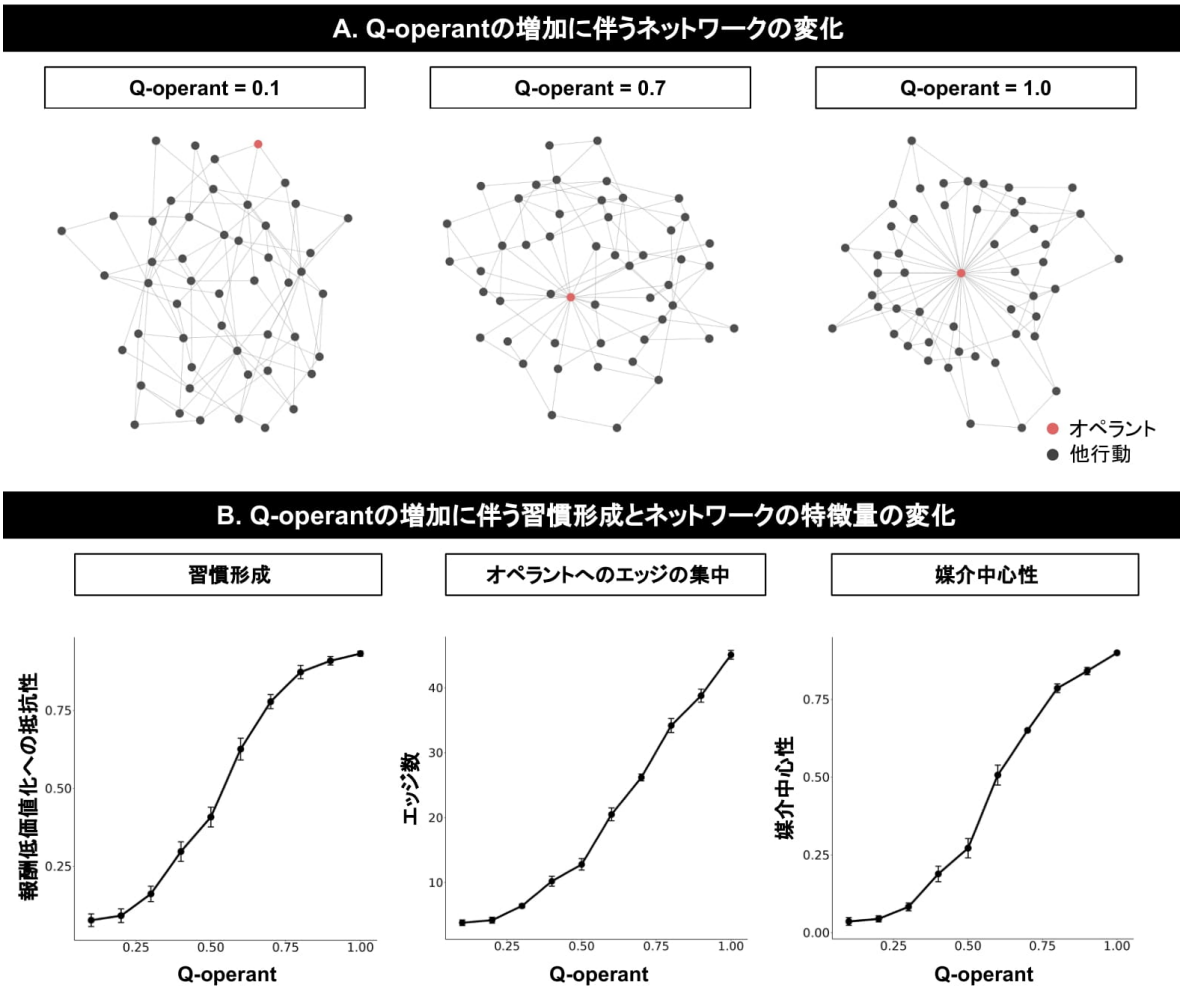

図 4.1.1. シミュレーション 1 の結果.

A. Q-operant の増加によるネットワーク変化を示す. 赤い点はオペラント反応, グレーの点は他行動をそれぞれ示す. B. Q-operant に対する習慣形成 (左), エッジの集中度合 (中央), そして媒介中心性 (右) を示す. 習慣形成の程度はベースラインの反応率に対する報酬低価値化後の反応率の比によって算出した. ベースラインからの減少が大きいほど, 小さい値となるため, 報酬低価値化に対する反応の抵抗性を示し, これが高いことは習慣形成が生じたことを意味する. オペラント反応へのエッジの集中は, オペラント反応が獲得したエッジ数によって評価した. 媒介中心性は任意の最短経路にオペラント反応が含まれる確率を示す. 折れ線の各点は 10 体のエージェントの平均値を示し, エラーバーはその標準誤差を示す.

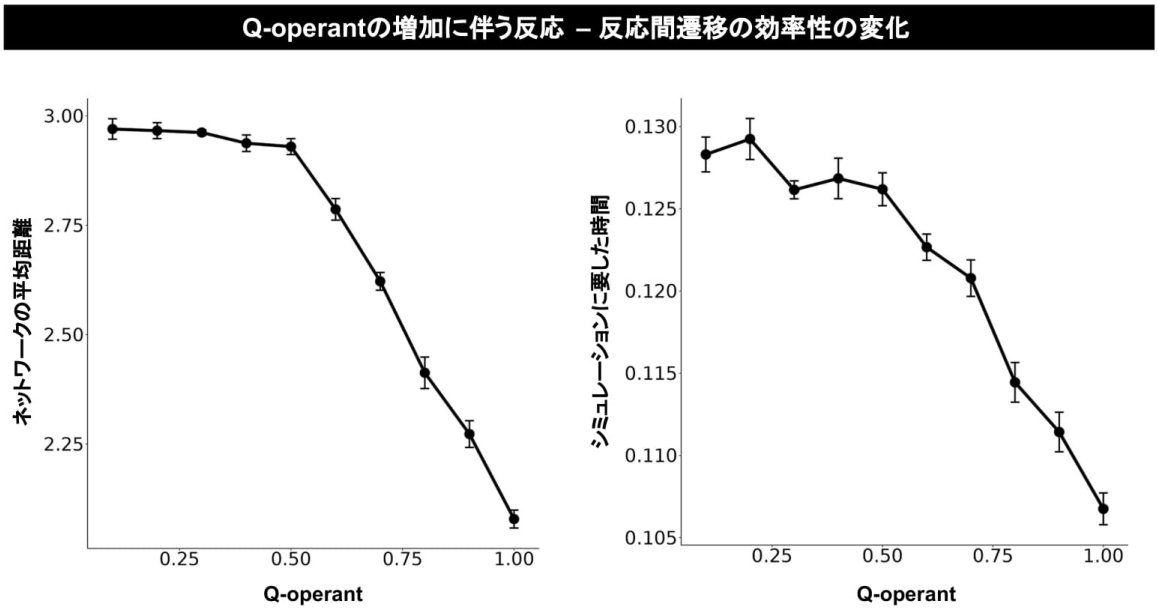

図 4.1.2. Q-operant の増加に伴うノード間遷移の効率性の変化.

左のパネルはネットワークの平均経路長, つまりネットワーク上の 2 つのノード間の最短経路の平均値を示す. 平均距離が短いほど, ある反応から別の反応への遷移が速くなる. 右パネルは, ベースラインのシミュレーションに要した時間を示している. 所要時間は実時間であり, シミュレーションの開始から終了までの時間である. ループの数は全てのシミュレーションで同じであるため, 所要時間の減少は, 最短経路探索と反応間の遷移の効率化を意味する. 折れ線の各点は 10 体のエージェントの平均値を示し, エラーバーはその標準誤差を示す.

実験 1 考察

実験 1 では疑似的にネットワークを生成することで, オペラント反応へのエッジの集中度合を操作して, 習慣形成が生じるかを検証した. Q-operant を系統的に操作することでオペラント反応へのエッジの集中度合を操作することに成功し, Q-operant の増加によってランダムネットワークから, オペラント反応のエッジの総取りが観察された (図 4.1.1. A, B の中央図). さらに Q-operant の増加によって報酬低価値化に対する抵抗性が上昇したことから (図 4.1.1. B の左図), オペラント反応へのエッジの集中によって習慣形成が生じたと結論付けることができる. テストフェイズではオペラント反応によって得られる報酬の価値は 0 としているため, 式 4.3 によってオペラント反応が選択されることはないにも関わらず, エージェントはオペラント反応に従事し続けた. これはオペラント反応が任意の 2 つの反応間を結ぶ最短経路に含まれるためである (図 4.1.1. B の右図). つまりエージェントはオペラント反応を選択しないものの, ある反応から別の反応へ遷移する過程でオペラント反応に従事した.

目的志向行動は, 環境に関する知識に基づいて柔軟に行動を決定する, 学習するモデルベース学習として捉えられ, 環境の知識に基づかないモデルフリー学習と比較して, 高い計算コストが要求される (Daw et al., 2005; Keramati et al., 2016). 習慣形成は, その意味において, 行動の柔軟性と引き換えに計算コストの削減として捉えることができる (Daw et al., 2005; Dezfouli and Balleine, 2012; Keramati et al., 2011). 実際に, 環境の知識に基づいたプランニングや学習には, 現在の状態から将来の状態を予測したり, 過去の経験を蓄積したりするような, 作業記憶の役割が重要となり (Daw et al., 2011; Keramati et al., 2016), ヒトを対象とした研究では, 実験参加者にストレスを与えることでモデルベースな計算を阻害することや, 習慣形成を促すことが報告されている (Otto et al., 2013; Schwabe and Wolf, 2009). 提案モデルにおいては, 作業記憶の役割を明示的に想定してはいないが, 最短経路探索による反応系列の生成は, 一種のプランニングとみなすことができる.

さらに, 習慣形成に伴って, 反応間の遷移の効率化やシミュレーションに要する時間が短縮したことは (図 4.1.2.), プランニングにおける計算コストの削減とみなすこともできる.

Garr et al., (2019) は二段階マルコフ決定課題で報酬低価値化手続きを行っており, そこでは反応列は獲得されたものの, 報酬への感受性は失われていないことを報告している. 本実験ではフリーオペラント事態を想定したシミュレーションだが, ネットワーク上の反応列生成でも, ネットワークの構造依存的に, 習慣形成が生じた. 式 4.3 でオペラント反応は選択されないため, エージェントは報酬感受性が欠如したわけではなく, 二段階マルコフ決定課題のように明示的な選択場面では報酬低価値の影響は受けることが予想される. この結果は Garr et al., (2019) の報告と合致し, De Houwer (2019) や Kruglanski and Szumowska (2020) の習慣が目的志向であるという指摘とも合致する.

実験 2：提案モデルによる習慣形成に関わる要因の効果の再現

実験 1 ではオペラント反応へのエッジの集中によって習慣形成が生じることが明らかになった. 実験 2 では, 既存の研究によって習慣形成を促す, あるいは阻害されるとされる要因が, 提案モデルに対して与える影響を検証する. そのため, 実験 2 では, 任意の実験環境下でエージェントに行動価値関数を学習させ, 習慣形成が生じるか検証する. フリーオペラント事態において, オペラント反応の習慣形成を促す, あるいは妨げる要因は 3 つある. 1 つ目は訓練量であり, ある状況下で 1 つの反応が繰り返し強化されると, その反応は習慣となる (Adams and Dickinson, 1981). 2 つ目の強化スケジュールである. VI スケジュールでは, 習慣形成が生じるが VR スケジュールでは阻害される (Dickinson et al., 1983). 第三の要因は選択肢の有無である. ある状況下で 2 つの選択肢があり, それぞれから異なる報酬が得られる場合 (例えば, 左レバー→餌, 右レバー→水), オペラント反応は習慣とならない (Colwill and Rescorla, 1985; Kosaki and Dickinson, 2010). ここでは, 上記の実験条件を再現し, これらの環境の下, 提案モデルで習慣形成が生じるかを検証する.

シミュレーション

実験 2 では, VI と VR スケジュール下で提案モデルを任意の報酬数で訓練することと, 2 つの選択肢が呈示される並立 VI VI スケジュール下での訓練を行った. VI と VR スケジュールでは強化率を統制するために, VR スケジュールで先に訓練を行い, そこで得られた報酬間間隔を記録して, VI スケジュールとして使用した. VR スケジュールと VI スケジュールではパラメータ設定は同じだが異なるエージェントを使用した. 訓練量は 25, 50, 100, 200 の 4 条件であり, エージェントが規定された報酬数を獲得したところで訓練を終了とした. ここでは VR 値を 15 に設定した. オペラント反応によって得られる報酬はシミュレーション 1 と同様に $\boldsymbol{r}_0 = 1.0$として他の要素については 0.001 とした.

並立 VI VI スケジュールではオペラント反応が 2 つ存在する. 基本的なシミュレーションの流れは単一スケジュールと同様であるが, 報酬の価値について 1 番目の反応をもう 1 つのオペラント反応とするため, $\mathbf{r}_0 = 1.0$, $\mathbf{r}_1 = 1.0$ として, 他の要素については 0.001 とした. さらに選択肢がない条件を設定し, そこでは並立 VI VT スケジュールを採用した. このスケジュールではオペラント反応に加えて, エージェントの反応に依存することなく時間経過によってランダムに報酬が呈示される. それによって単一スケジュールでも強化率を並立 VI VI スケジュールと統制して, 選択肢の有無の影響を検討できる. ここで使用したスケジュール値は並立 VI 60 VI (VT) 60 でありエージェントが 200 回報酬を獲得したところで, 訓練を終了した. 全ての実験で他行動のスケジュールを FR 1 とした.

訓練が終了した後に式 4.2 に従ってネットワークを生成して, ベースラインとテストフェイズを行った. 基本的な流れは実験 1 と同様であるが, 並立スケジュール事態においては一方の報酬価値のみを低下させるため, $\mathbf{r}_0 = 0$としてもう一方の報酬価値は操作しなかった.

結果

VI と VR の両スケジュールで訓練量の増加に伴って, オペラント反応は報酬低価値化への抵抗性を増加させ, 他のノードからのエッジを多く獲得するようになり, 媒介中心性が上昇した (図 4.2.1. A). VI と VR を比較すると, 全ての指標で VI が VR を上回った (図 4.2.1. A). 選択肢の有無を比較すると, 選択肢が存在する事態 (並立 VI VI) では, 報酬低価値化への抵抗性が低いのに対して, 選択肢が存在しない事態 (並立 VI VT) では抵抗性が高かった (図 4.2.1. B の左図). それぞれのネットワークを比較すると, 選択肢がある事態では 2 つのオペラント反応が同程度に, ネットワーク上の殆どのノードを獲得したが, 選択肢がない事態では, オペラント反応のみが多くのエッジを獲得した (図 4.2.1. B の中央・右図). VI と VR スケジュール下でのオペラント反応の自己遷移が有する行動価値関数は訓練量に伴い上昇し, VR スケジュールが VI スケジュールを上回っていた (図 4.2.2.).

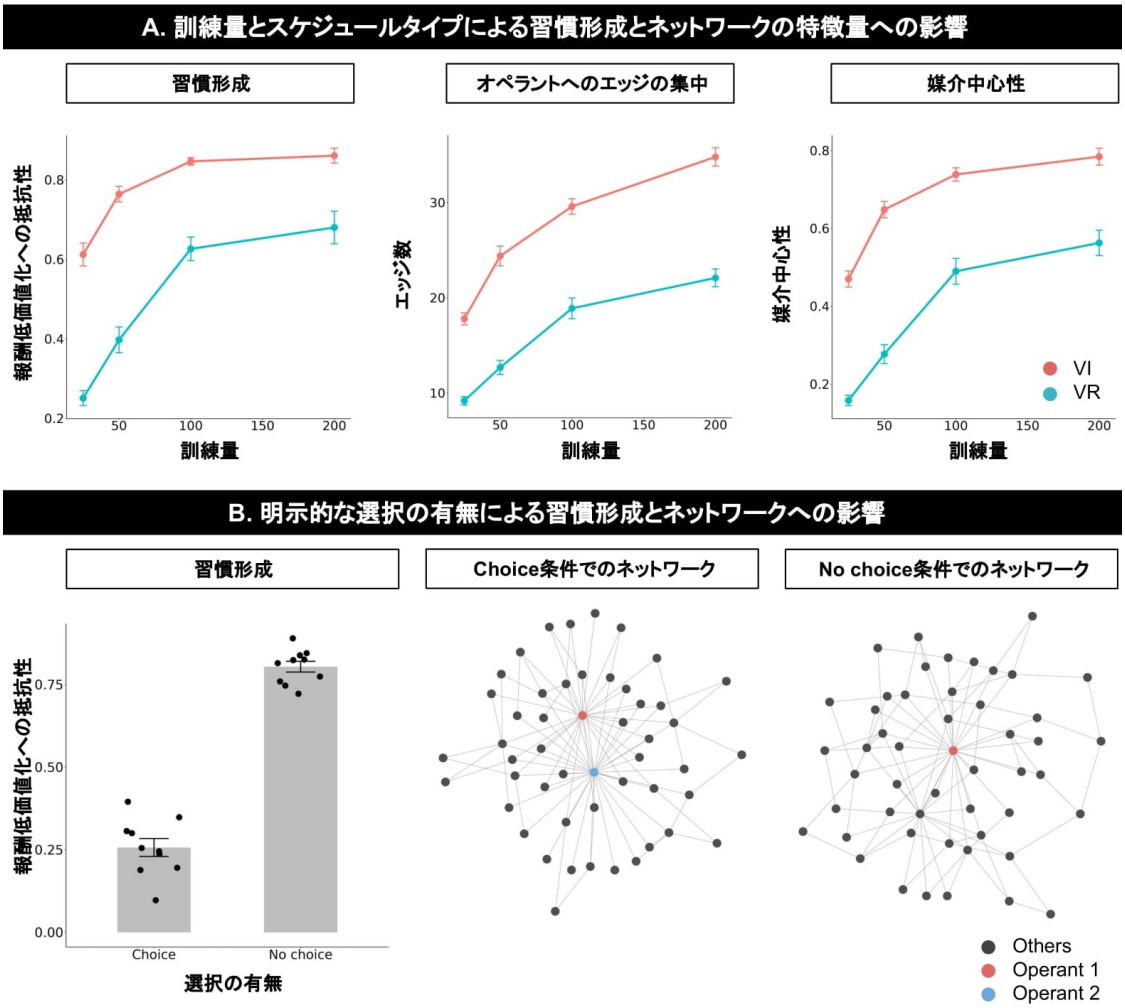

図 4.2.1 シミュレーション 2 の結果.

A. 訓練量とスケジュールタイプが習慣形成とオペラント反応へのエッジの集中度合いと媒介中心性に与える影響. 赤と青の線はそれぞれ, VI と VR スケジュールを示し, 訓練量の増加に伴う各指標 (左: 報酬低価値化への抵抗性, 中央: オペラント反応へのエッジの集中度合, 右: 媒介中心性) を示す. B. 選択肢の有無が習慣形成とネットワークに対して与える影響. 左パネルは報酬低価値化への抵抗性を示す. 中央のパネルは選択がある事態でのネットワークの構造を示し, 赤と青がそれぞれの選択肢に対応したオペラント反応を示す. 右のパネルは選択肢が存在しない事態でのネットワーク構造を示す. 折れ線の各点と棒グラフの各バーは 10 体のエージェントの平均値を示し, エラーバーはその標準誤差を示す.

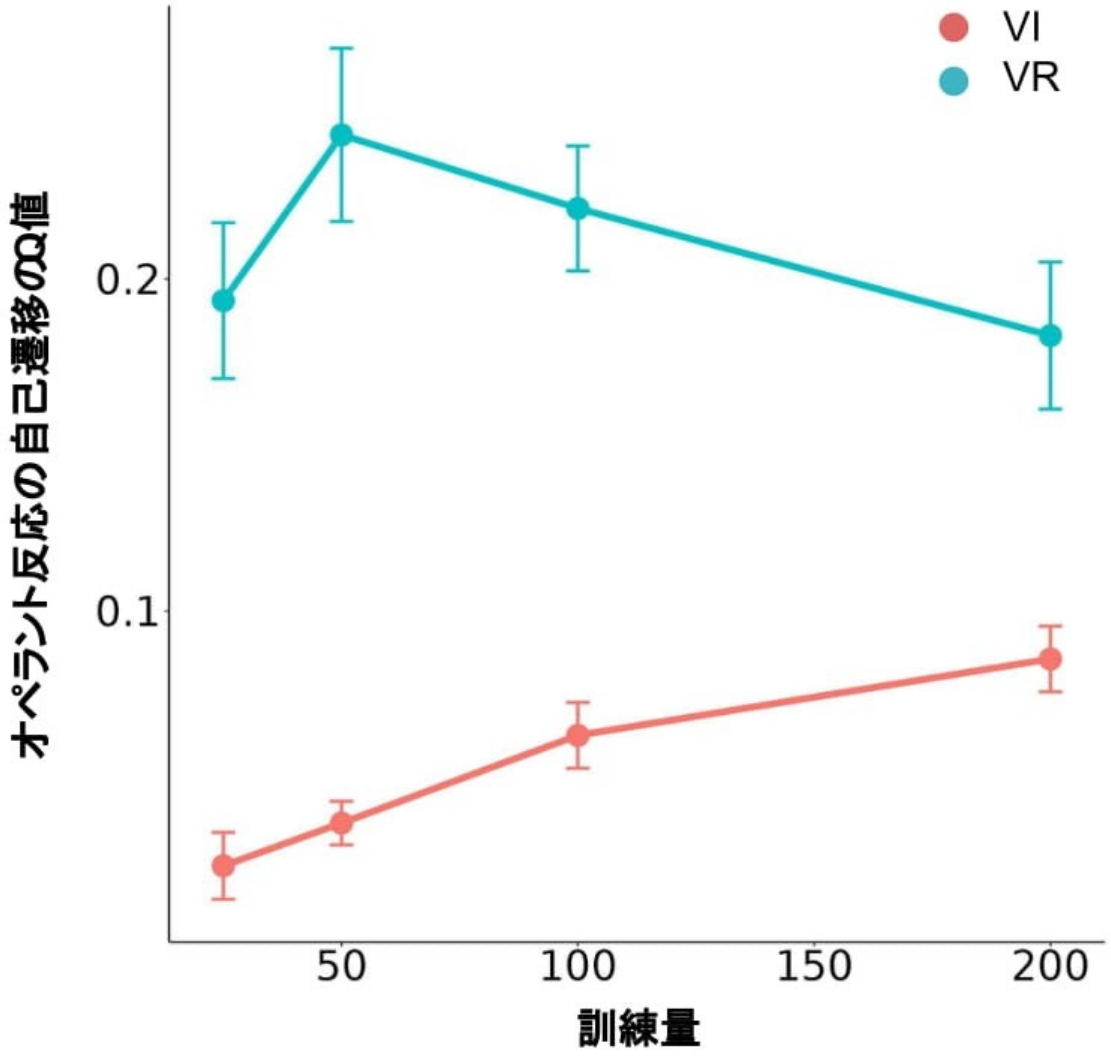

図 4.2.2. 訓練量に対するスケジュールごとのオペラント反応の自己遷移の行動価値関数の変化.

折れ線の各点は 10 体のエージェントの平均値を示し, エラーバーはその標準誤差を示す.

実験 2 考察

実験 2 では訓練量, スケジュールタイプ, そして選択の有無が習慣形成に与える影響を検討した. 過去の実験的な研究によって, 訓練量の増加による習慣形成の促進, VR スケジュールと選択事態では, 習慣形成の阻害が報告されている (Adams and Dickinson, 1981; Colwill and Rescorla, 1985; Dickinson et al., 1983; Kosaki and Dickinson, 2010). これらの結果は全て提案モデルによって再現され (図 4.2.1.) , 実験 1 で示したように, それらはエッジのオペラント反応への集中によって説明できる. しかし, 選択事態では双方のオペラント反応が同程度にエッジを獲得しているにも関わらず, 報酬が低価値化された反応のみが減少した (図 4.2.1. B の左図). これは, モデルでは反応を報酬の価値に基づいて選択していることで, ベースラインと比較して反応が選択されることがなくなったこと, そしてもう一方のオペラント反応もエッジを多く獲得したことで, 最短経路が複数存在することに起因する.

訓練量の増加と VI による習慣形成は実験 2 で示したようにオペラント反応へのエッジの集中によって説明できる. 訓練量の増加に伴ってオペラント反応が獲得したエッジ数が増加しており, 特にそれは VI で顕著であった (図 4.2.1. A の中央図). VI スケジュールでは強化子が時間経過に依存して呈示されるため, 前の反応からの経過時間が長くなるにつれて強化確率が上昇する. 従って他行動からオペラント反応への遷移が分化強化されることで, 他行動からオペラント反応への遷移の行動価値関数が高くなり, エッジがオペラント反応に集中した. VR スケジュールでは強化確率は前の反応からの経過時間には依存しないため, VI と異なり連続した反応が強化されやすい. VR スケジュールでオペラント反応の自己遷移の行動価値関数が, VI スケジュールの自己遷移の行動価値関数を上回っていることは (図 4.2.2.), 上記のようなスケジュール特性によるものである. つまり, VI 及び VR スケジュールでの習慣形成の程度の差は, スケジュールの持つ経過時間と強化確率の性質によって説明できる.

実験 3：モデルによる習慣形成を規定する要因の検討

実験 2 では過去の研究で報告されてきた習慣形成を促す, あるいは促進する要因が提案モデルに及ぼす影響を検討し, 実際の動物実験と同様の結果を再現することに成功した. 提案モデルではオペラント反応へのエッジの集中によって習慣形成が生じるとしており, これはシミュレーション 1, 2 の結果から支持される. しかし, そうしたネットワークの構造が環境のどのような要因に起因しているかは明らかではない. 実験 3 では習慣形成を促す原因となる要因をシミュレーションによって明らかにする.

従来の習慣形成に関する理論では反応-報酬の相関関係によって規定されるとされている (Dickinson, 1985). 具体的には反応と報酬の間の相関性が知覚される場合には習慣形成が阻害され, この相関性が知覚されなくなることで習慣形成が生じるとされている. VR スケジュールでは反応率に対して線形に強化率が上昇するのに対して, VI スケジュールではある一定の反応率以上は殆ど, 強化率に影響を及ぼさない (図 4.3.1. 左図). 従って VR スケジュールにおいては習慣形成が阻害されるが, VI スケジュールでは習慣形成が生じる.

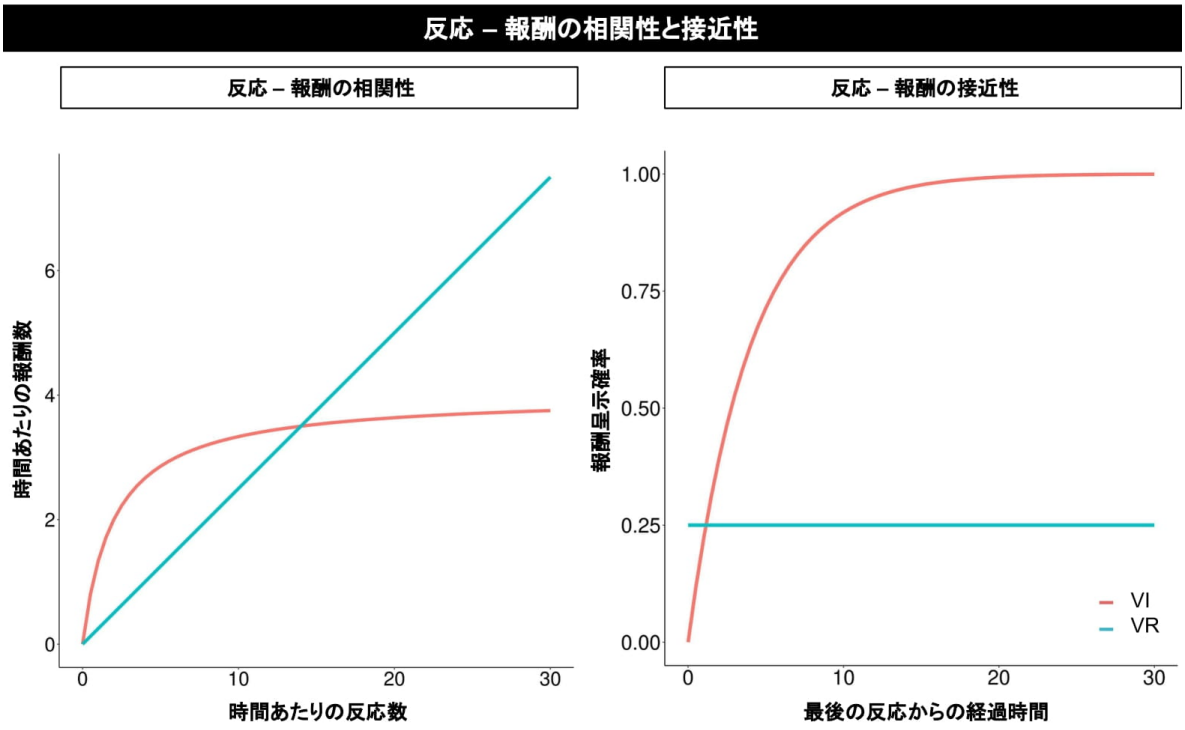

図 4.3.1 VI と VR スケジュールの巨視的・微視的なスケジュール特性

左は VI と VR の巨視的なスケジュール特性 (反応-報酬の相関性) を示す. VR では反応に対して報酬率は線形に上昇するが, VI では負の加速度的な関数となり, 一定以上の反応は報酬率の上昇に寄与しない. 右は VI と VR の微視的なスケジュール特性(反応-報酬の接近性)を示す. VI スケジュールでは経過時間に伴って報酬確率が上昇するが, VR スケジュールでは前の反応からの経過時間に関わらず報酬確率は一定であるため, VR では反応と報酬の接近性が高い.

一方で近年の研究では, 反応-報酬の相関性ではなく, 反応-報酬の接近性によって習慣形成が規定される可能性が報告されている (Corbit et al., 2014; De Russo et al., 2010; Garr et al., 2020). 例えば, De Russo et al. (2010) は VI および FI スケジュールでマウスを訓練した. FI と VI は反応率と強化率の相関関係は同じであり, 両スケジュールとも動物がいくらオペラント反応に従事しても, 一定時間内に決められた数以上の報酬を得ることができない. このような条件下では, 反応-報酬相関説は両スケジュールが同程度の習慣形成を導くと予測する. しかし, 実際には FI スケジュールで訓練されたマウスのオペラント反応は目標指向的であったが, VI スケジュールではオペラント反応は習慣となっていた. DeRusso et al. (2010) は, 個々の反応と報酬の平均時間距離で定義される接近性が習慣形成を阻害すると結論づけている. FI スケジュールでは, 動物は報酬が呈示される時間に近づくにつれ, より多くの反応を自発する傾向がある. 一方, VI スケジュールでは, 動物はいつ報酬が得られるかわからないため, 報酬間の間隔が均一な反応をする. このように, FI スケジュールでは, 動物は報酬の直前に多くの反応を自発するため, 反応と報酬の接近性が高くなるが, VI スケジュールではオペラント反応は一様に分布するため接近性は低くなる.

同様に VI と VR スケジュールでは VR スケジュールで接近性が高くなる. これはスケジュールによって規定される報酬確率の時間的な変化に依存する. VR スケジュールは反応依存のスケジュールのため, 時間経過に関わらず反応によって得られる報酬の確率は一定だが, VI スケジュールでは時間経過に伴って上昇する (図 4.3.1. の右図). これは VI スケジュールにおいてはバウト内の反応のような短い IRT の反応には報酬が与えられず, 中-長程度の IRT に選択的に報酬が与えられることを意味する. 従って VI スケジュールでは反応-報酬接近性が低くなり, 習慣形成が促進されることとなる.

この議論は行動分析学における VI-VR 反応率差の議論と類似している. VI と VR スケジュールは時間的に類似した反応パターンを生成するが, 全体的な反応率は

VR スケジュールが上回る (Baum, 1993; Baum and Grace, 2020; Ferster and Skinner, 1957). この反応率差には 2 通りの説明がある. 第一の説明は, 反応-報酬の巨視的な関係に基づくものである (Baum, 1973, 1981). VR スケジュールでは, 動物がより多くの反応を示すほど, より多くの報酬を得ることができる一方, VI スケジュールでは, 動物がどのように反応しても実験的に定義された以上の報酬を得ることはできない (図 4.3.1 左実線). このような反応と報酬の相関性を動物が知覚することで, VR でより高い反応率になると考えられている. この説明は, 習慣形成における反応-報酬相関説 (Dickinson, 1985; Perez and Dickinson, 2020) と同様のものである. もう 1 つは強化子が呈示される直前の IRT による説明である (Wearden and Clark, 1988; Tanno and Silberberg, 2012). VI スケジュールでは, 最後の反応からの経過時間が長いほど強化子を得られる確率が高くなるため, 長い IRT が強化されやすい. 一方, VR スケジュールでは直前の反応の経過時間に関わらず強化確率は一定である(図 4.3.1. の右図). さらに反応がバウト・休止パターンという構造を持ち, VR ではバウト長が VI より長いため, 全反応に占めるバウト内反応が多い. 従って VR では長い IRT よりバウト内の短い反応が強化されることが多くなる. この IRT への分化強化によって VI と VR の反応率差が生じる (Wearden and Clark, 1988; Tanno and Silberberg, 2012). 実験 2 の図 4.2.2.でも示されたように, VI では自己遷移の行動価値関数がごくわずかであるのに対して, VR では自己遷移の行動価値関数が比較的高い値となっているように, 提案モデルはスケジュールの反応-報酬接近性の特性を反映している. 従って提案モデルは Perez and Dickinson (2020) とは対照的に, IRT 強化と同様のスケジュールの接近性による説明を採用する.

Peele et al. (1984) は VI-VR 反応率差が生じる要因を実験的に明らかにした研究である. ここでは, スケジュールの相関性と接近性を操作する手続きが用いられた. 使用されたのは tandem VI VR と tandem VR VI スケジュールである. Tandem VI VR スケジュールは VI スケジュールをスケジュールの中核に据えることで, 反

応と強化子の間の相関性を小さくしたが, VI 終了後の微小な VR スケジュールによって短い期間での連続した反応でも強化されやすくなる. つまり巨視的には VI スケジュールの特性を持ちながらも, 微視的には VR スケジュールの特性を持つスケジュールである. Tandem VR VI スケジュールは VR スケジュールをスケジュールの中核に据えることで, 反応と強化子の間の相関性を高くするが, VR 終了後の微小な VIR スケジュールによって比較的長い IRT が強化されやすくする. つまり巨視的には VR スケジュールの特性を持ちながらも, 微視的には VI スケジュールの特性を持つスケジュールである. 実験 3 では上記のスケジュール下での提案モデルの振る舞いを分析することで, 提案モデルが 2 つの習慣形成の仮説のどちらを支持するか検証し, 上記の手続きが, 習慣形成に関する 2 つの仮説を検証する実験としての有用性を検証する.

シミュレーション

実験 3 では単一の VI, VR スケジュールに加えて, tandem VI VR スケジュールと tandem VR VI スケジュールで実験を行う. tandem VI VR スケジュールは VI スケジュールと同様の反応-報酬の相関関係を持つ一方で, 時間経過に伴う報酬確率の変化については VR スケジュールと同様の性質を持つ. Tandem VR VI はその逆に, 反応-報酬の相関関係は VR スケジュールに共通するものの, 時間経過に伴う報酬確率の変化については VI スケジュールと同様の性質を持つ. シミュレーションでは tandem VI 15 VR 3 と tandem VR 10 VI 5 を採用した. 単一の VI では tandem VI VR 3 の報酬間間隔を用いて強化率を統制した. 単一の VR スケジュールでは tandem VR 10 VI 5 の反応数を用いて反応当たりの強化確率を統制した.

結果

反応-強化子の相関性と接近性がモデル対して及ぼす影響を検討するために, 相関性と接近性を操作したスケジュール下での習慣形成とネットワークの特徴量を検討した. 単一 VI と tandem VI VR は両スケジュールでも反応 – 強化子の相関性

は低いが, 反応 – 強化子の接近性は, tandem VI VR が高くなる. 一方で単一 VR と tandem VR VI は,相関性は高いが, tandem VR VI では接近性が小さくなる. これらのスケジュールでは, 接近性の低い単一 VI, tandem VR VI で報酬低価値化への抵抗性, オペラント反応が獲得したエッジ数, そして媒介中心性が単一 VR と Tandem VI VR を上回った (図 4.3.2.). オペラント反応の自己遷移が有する行動価値関数では, 接近性が高い単一 VR と tandem VI VR が単一 VI と tandem VR VI が上回った (図 4.3.3.).

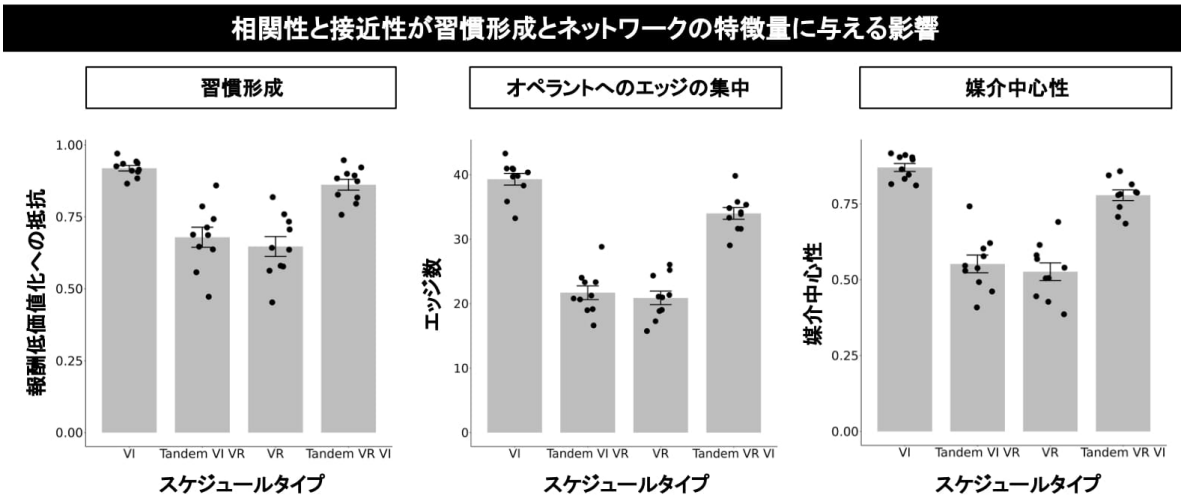

図 4.3.2. シミュレーション 3 の結果.

VI, tandem VI VR, VR, tandem VR VI の各スケジュールでの報酬低価値化 (左), オペラント反応へのエッジの集中 (中央), そして媒介中心性 (右) を示す. 棒グラフの各バーは 10 体のエージェントの平均値を示し, エラーバーはその標準誤差を示す.

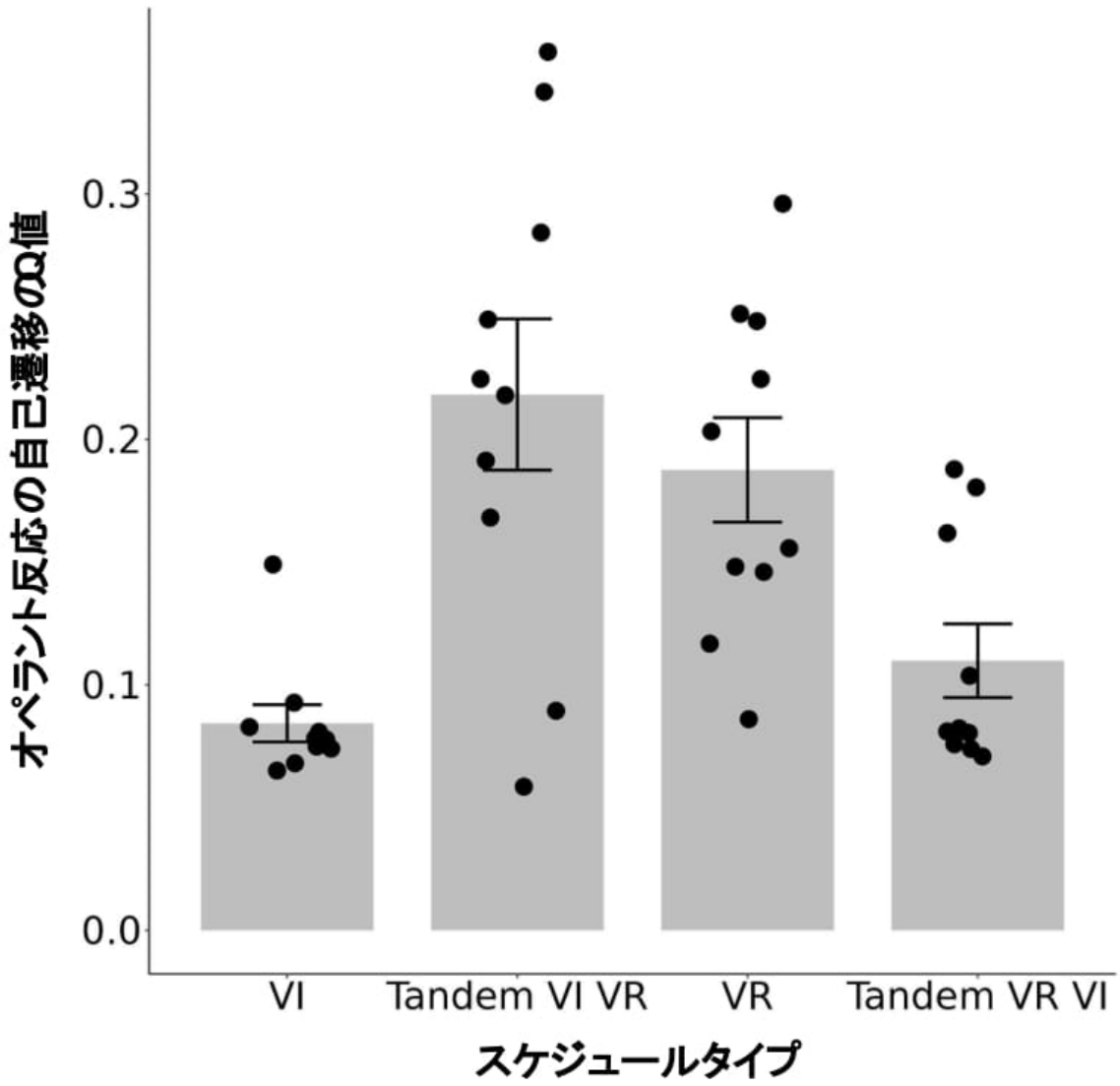

図 4.3.3.スケジュールごとのオペラント反応の自己遷移の行動価値関数.
棒グラフの各バーは 10 体のエージェントの平均値を示し, エラーバーはその標準誤差を示す.

実験 3 考察

実験 3 では通常の VI と VR スケジュールに加えて, 巨視的には VI, 微視的には VR の特性を持つ tandem VI VR スケジュールと, 巨視的には VR, 微視的には VI の特性を持つ tandem VR VI スケジュールで習慣形成が生じるか検討した. 従来の反応-報酬相関説では, スケジュールの巨視的な特性によって習慣形成が生じるとされ, 反応と報酬の相関性が低い VI スケジュールで習慣形成が生じて, 相関性の高い VR スケジュールでは習慣形成が阻害される, と説明されてきた. 本実験の結果で採用した単一 VI と tandem VI VR は双方とも巨視的には VI の特性を持ち, 単一 VR と tandem VR VI は巨視的には VR の特性を持つ. 従って反応-報酬相関説によれば, 単一 VI と tandem VI VR で習慣形成が生じて, 単一 VR と tandem VR VI では習慣形成が阻害されると予測される. 予測に反して, tandem VI VR は巨視的に VI の特性を持つにも関わらず単一 VI に比較して習慣形成が阻害され, 巨視的に VR の特性もつ tandem VR VI では単一 VR と比較して習慣形成が促された (図 4.3.2). 習慣形成が生じた単一 VI と tandem VR VI は微視的には VI スケジュールの特性を持ち, 習慣形成が阻害された単一 VR と tandem VI VR は微視的には VR スケジュールの特性を持つ. 従って提案モデルではスケジュールの微視的な特性によって習慣形成が生じていると結論づけることができる. これは近年の反応-報酬接近説による説明に立つことを意味する (Corbit et al., 2014; De Russo, et al., 2010; Garr , et al., 2020).

実験 3 では提案モデルが VI と VR の習慣形成の差を説明する方法が, 行動分析学における VI-VR 反応率差の IRT 強化と類似していることに着目して, VI-VR 反応率差の議論で決定的な実験となった Peele et al. (1984) で使用されたスケジュールを採用した. Peele et al. (1984) は tandem VI VR と tandem VR VI スケジュールによって VI-VR 反応率差がスケジュールの微視的な特性によって生じていることを示した. つまり, Perez and Dickinson (2020) は強化学習的なモデルによって Baum (1973, 1981) で提案された反応-報酬相関説を実装したのに対して, 本実

験は強化学習モデルによって IRT 分化強化と同様の説明様式を取り入れて習慣形成の説明に成功した. 本シミュレーションで採用した, Peele et al. (1984) で使用したスケジュールで, 習慣形成が生じるか検討することは, 習慣形成における心理学的な要因を明らかにする上で有用だろう.

研究 4 総合考察

本研究では, 動物の行動を反応のネットワークと見なし, 習慣形成をネットワーク構造の変化として説明した. 実験 1 では, 任意のネットワークを生成することで, ネットワーク構造の違いによる報酬低価値化の影響を検討した. その結果, オペラント反応がネットワーク上のエッジの大半を獲得することで習慣形成が生じることを明らかにした. 実験 2 では, 先行研究で, 習慣形成を促進または阻害すると報告されている環境を模して, 提案モデルでも習慣形成が生じるか検討した. シミュレーション結果は先行研究の結果と一致し, 提案モデルの行動レベルでの妥当性が示唆された. 実験 3 では, 習慣形成の心理学的な仮説である反応-報酬相関説と反応-報酬接近説の内, 提案モデルはどちらの説を支持するか検討した. 提案モデルは反応-報酬接近説を支持すると同時に, シミュレーションで用いた実験事態がこの 2 つの仮説を検証する上での有用性を示した.

既存の理論の多くでは, 習慣形成を 2 つの異なるシステムの競合として捉えている (Daw et al., 2005; Perez and Dickinson, 2020). そこでは目的志向システムと習慣システムの 2 つが仮定されており, 1 つの反応に対するシステムの支配性によって目的志向行動か習慣か決定される. 提案モデルでは, 行動をより巨視的な時間スケールで捉えることで新たな説明を与えることに成功した. 提案モデルは既存のモデルと大きく異なる点が 2 つある. 第一に, 行動を単一の要素としてではなく, 相互に接続された反応のネットワークとして捉えることである. 従来のモデルでは, レバーを押す, ボタンを押すといった実験事態での反応にのみ注目しており, 実際の生物が従事しうる他の行動を無視していた. ヒトを含む動物の反応は, 互いに独立したものではなく, 前後の反応と関連づけられている. 提案モデルでは, こ

のような反応の関係をネットワークとして表現し, 習慣形成をその構造の変化として説明した. 第二に, 提案モデルでは, 目標指向行動と習慣を競合的に扱わず, Dezfouli and Balleine (2012) で提案されたような階層的な制御関係を想定している. モデルでは, 全ての反応は報酬の値に基づく選択と, ネットワークにおける最短経路の探索によって決定されると仮定している. つまり, すべての行動は, シミュレーションで示した報酬感受性の有無によらず, 常に目標指向的である. しかし, ネットワークの構造によって, 任意の 2 つの反応間で可能な経路が制約され, 結果として習慣が形成されることとなる. こうした習慣形成の新たな捉え方は, ヒトを対象とした 2 段階マルコフ決定課題に限られていたが, 本研究ではフリーオペラント事態における, 報酬価値感受性の欠如をも説明できることを示した.

提案モデルは, ラットを用いたフリーオペラント事態における習慣形成に関する実験を扱っている (Adams, 1982; Dickinson et al., 1983; Kosaki and Dickinson, 2010) , ここで議論した実験のほとんどは Perez and Dickison (2020) でも扱われており, 提案モデルでも Perez and Dickison (2020) のモデルでも実験結果と整合的な結果を再現している. Perez and Dickison (2020) では報酬反応相関に基づく説明をしているが, 提案モデルでは反応-報酬接近性に基づく説明を採用している (DeRusso, et al., 2010; Garr, et al., 2020). シミュレーションではこの 2 つの仮説が異なる結果を予測する実験として Peele et al.(1984) の手続きを採用したが, 実際の動物実験では, 未だに検討されていないため, 今後の習慣形成の研究で有用な知見をもたらすものと期待される.

提案モデルは Dezfouli and Balleine (2012) と 2 つの類似性がある. 彼らのモデルでは他行動は仮定していないものの, 習慣形成を反応列の獲得によって説明している. 提案モデルでは反応列を直接学習することはないが, 最短経路探索によって反応列の生成を行っている. 第二に彼らのモデルでも反応列の生成の前段階に目的志向的な選択という過程が組み込まれている. これらは提案モデルにおける報酬の価値に基づいた選択に類似している. しかし Dezfouli and Balleine (2012,

2013, 2014) の一連の研究では主にヒトの二段階マルコフ決定課題を対象に実験, シミュレーションが行われている. フリーオペラント事態への適用は訓練量の習慣形成への影響に留まり, 他のスケジュールの種類や選択肢の有無の影響は検討されていない. さらに Garr et al. (2020) では反応列の獲得とフリーオペラント事態の報酬感受性の欠如による習慣の定義が異なる可能性が支持されたことから, Dezfouli and Balleine (2012) で提案された反応列の生成というアイデアのフリーオペラント事態への適用可能性が疑わしいものとなった. しかし, 提案モデルは, 彼らのモデルと共通の仮定を持ちながらも, フリーオペラント事態での結果を再現することができた (Adams, 1982; Colwill and Rescorla, 1985; Dickinson, et al., 1995; Dickinson et al., 1983; Kosaki and Dickinson, 2011). 従って, 二段階マルコフ決定課題とフリーオペラント事態での習慣形成は現象として共通のものであるかは自明ではなかったが, 本研究によって二段階マルコフ決定課題で提案された反応列の獲得と類似した方法で報酬感受性の欠如を説明することに成功した.

習慣の形成や反応列の生成には皮質・線条体ネットワークが関与している (Graybiel, 1998; 2008). 特に, 背外側線条体 (DLS)は, 目標指向行動から習慣への移行に重要であることが知られている (Yin et al., 2004). DLS の活動は訓練による習慣形成にともない変化することが報告されており (O'Hare et al., 2016; Tang et al., 2007), 習慣形成後の DLS の損傷によって習慣が目標指向行動へと戻る (Yin et al., 2006). また, DLS は反応列の生成 (Yin, 2010)や運動ルーチンの形成 (Jurado-Parras et al., 2020) も担っている. 学習性行動だけでなく, DLS は生得的な反応列もコードしている (Aldridge and Berridge, 1998). これらの事実は, 習慣形成と反応列の生成には共通の神経基盤があることを示唆している.

最近の研究では, DLS は反応列に関する情報だけでなく, 反応のトポグラフィに基づいて分類された反応とその間の遷移も符号化することが報告された (Markowitz, et al., 2018). 彼らは, オープンフィールドの状況下でマウスの DLS 活動をファイバーフォトメトリーで記録し, 各反応と相関のする神経活動を報告し

た. さらにその活動は前後の反応によって異なり, DLS は反応間の遷移も符号化していた. これらの結果は DLS 反応間の遷移といったネットワーク的な情報表現をしていることを示唆する. さらに, DLS を損傷させたマウスの反応の遷移確率行列は, よりランダムな遷移を示した. 提案モデルでランダムネットワークからオペラント反応へのエッジの集中によって習慣形成が生じることを示したが, Yin et al. (2006) で DLS 損傷によって習慣が目的志向行動へ戻ったことを, 行動ネットワークのエントロピーの上昇によって説明できる可能性を示唆する.

前頭前野 (prefrontal cortex; PFC), 背内側線条体 (dorsomedial striatum; DMS), 腹側線条体 (ventral striatum; VS)からなる Corticostriatal circuits は目標指向行動に関与する (Balleine and O'Doherty, 2010). DMS は, 目標指向行動の獲得, 結果に対する感受性の維持, 目標指向行動の発現に関与することが知られている (Ostlund and Balleine 2005, Yin et al., 2005). DMS は PFC から興奮性の入力を受け, DLS は感覚運動野と運動前野から入力を受ける (Yin et al., 2005). 習慣形成の二分法的な見方では, 目標指向的な行動は, 訓練後に習慣に置き換えられる. このような場合, DLS の寄与は DMS の寄与よりも大きくなる (Yin et al., 2004；Yin et al., 2005). しかし, 訓練を行っても, PFC, 前帯状皮質 (anterior cingulate cortex), VS, DMS など多くの脳領域が報酬の予期によって調節される (Niki and Watanabe, 1979; Schultz et al., 1992; Shidara and Richmond, 2002; Watanabe, 1996; Yin et al., 2004; Yin et al., 2005). 提案モデルでは, エージェントのあらゆる反応は目標指向であると仮定しているため, 訓練の段階に関わらず, エージェントは報酬の値に基づいて応答を選択する. 従って目標指向行動を含む領域は, 訓練を行っても報酬の期待によって調節されるという事実は, 提案モデルの仮定では矛盾なく解釈することができる.

提案モデルには 2 つの大きな限界がある. 第一に, 提案モデルは現実の生物が持つ制約を考慮していない. 例えば学習における本能的逸脱 (Breland and Breland, 1961) やスケジュール誘導性行動 (Falk, 1966; Hymowitz, 1971; Gentry, 1968;

Levitsky and Collier, 1968) の存在は, 任意の反応間にはエッジが貼られやすい, あるいは生得的に存在することを示唆する. さらには, 身体的な制約によって不可能な反応遷移なども存在する. 提案モデルでは全ての反応を一様に扱っているためこれらの事実を扱っていない. こうした生得的な制約をエッジの貼られる確率の事前分布として表現することで問題を解決できると考えられる. それによって, 従来の学習心理学や行動分析学で理論的に扱われることが少なかった現象についても, ネットワークという視点から説明できる可能性がある.

第二に, 我々のモデルでは, フリーオペラント事態での習慣形成の実験にしか対応できない. 実験で行ったシミュレーションはすべてフリーオペラント事態での実験であり, ヒトの二段階マルコフ決定課題での実験ではない. これは我々のモデルに特有の問題ではなく, 他の既存モデルもどちらか一方を扱っているか, もしくは, 扱っていたとしてもごくわずかである. 従って, 両実験系で多くの実験が行われているが, それらの間の手順の違いや結果の同一性については体系的に検討されていない. 習慣形成についてより統一的な理解を得るためには, 既存の研究で採用され, 得られた手順や結果を系統的に分析する必要がある.

研究 4 では, 研究 3 で導入した他行動という視点を, 行動の長い時間スケールで捉えた時に現れる, 反応が相互結合したネットワークとして捉えることで, その構造変化によって習慣形成が生じるという, 従来とは異なる習慣形成のモデルを提示した. 提案モデルは既存の実験結果に新しい解釈を与えるだけでなく, 習慣形成とは異なる文脈の, VI-VR 反応率差やバウト研究との関連性を明らかにすることで, VI-VR 反応率差における実験事態が, 習慣形成に関する 2 つの理論が相互排他的な予測をもたらす実験事態を提案することにも成功した. また, 行動をネットワークとして捉えることで, フリーオペラント事態における習慣形成の神経基盤と, 二段階マルコフ決定課題における反応連鎖に基づいたモデルや, その神経基盤との関わりについても解釈することが可能となった. 研究 3 と 4 で取り入れた他行動は, 実際に計測することの困難さからか, 従来の行動分析学と学習心理学では取

り扱われてこなかった. しかし, 現在では網羅的な行動計測が可能となっており (Markowitz et al., 2018; Wiltschko et al., 2020),それによってもたらされる大規模な行動データを解釈する上で, この他行動や, そのネットワーク構造という視点は重要な意味を持つことになるだろう.

研究 1−4 まとめ

研究 1 から 4 を通して, 行動分析学と学習心理学で扱われてきた様々な現象を, 強化学習の枠組みでモデル化することで, パブロフ型条件づけとオペラント条件づけ, 及びそれらが関わる行動現象を, 強化学習によって包括的に扱えることを示した. 強化学習という理論的な側面からのアプローチに加えて, 深層学習や画像処理による計測系の改良は, 新たな指標の定量化が可能とし, 動物の学習における新たな側面を明らかにすることに成功した.

研究 1 では, パブロフ型条件づけ事態において, 瞳孔サイズが報酬に対する予測を反映することを明らかにした. 従来の実験装置では, マウスが実験中でも自由に動き回ることができたため, その瞳孔サイズの計測は困難であったが, 頭部固定装置と深層学習による画像解析によって, 課題中の瞳孔サイズの定量化に成功した. さらに, マウスの瞬間的な反応の動態を強化学習モデルによって記述することで, 課題中のマウスの報酬予測や報酬予測誤差の動態を推定し, 瞳孔サイズの動態を報酬予測の動態から予測することに成功した. さらに, この瞳孔サイズの報酬予測的な変化は, 薬理処置によってリッキングを抑制しても生じた. これらの結果から, 瞳孔サイズが報酬予測を反映することが明らかになった. このように, 強化学習によって瞬間的な反応の動態をモデル化することを通して, 動物の内的な報酬予測の動態を明らかにしたが, これは強化学習が行動のモデル化のツールとしてだけではなく, 行動データから, 動物の潜在的な過程を推定するツールとしての有用性を示すものである. さらに, 深層学習と画像処理を駆使することで, 定量化に成功した瞳孔サイズという指標と結びつけて解析することで, 動物の学習や予測に関わるメカニズムに迫ることが可能となった.

研究 2 では, オペラント反応の瞬間的な動態を, 強化学習によってモデル化することで, 報酬確率が消去バーストの制御要因であることをシミュレーションによって同定し, マウスの実験でそのモデルの予測を検証することで, 消去バーストが

報酬確率によって制御されていることを明らかにした. 従来の消去を扱う理論では, 行動や行動を制御する潜在過程の単調減少として消去を捉えていたため, 消去によって生じる一過性の反応の上昇は説明できない (Esber and Haselgrove, 2011; Mackintosh, 1975; Nevin and Grace, 2000; Pearce and Hall, 1980; Pearce et al., 1982; Rescorla and Wagner, 1972). そこで, 消去を単調減少する単一の過程ではなく, 強化学習における好奇心というアイデアに着目することで, 消去の開始直後に一時的に上昇するもう 1 つの潜在過程を導入, それら 2 つの過程の和として, 消去という現象を捉え直した. 好奇心駆動型強化学習という消去の新たな側面に光を当てたモデルは, 従来の理論では説明しえない消去バーストが生じることを予測し, シミュレーションによって, 消去バーストの制御要因を同定することに貢献した. このモデルによる予測を裏付けるために, 頭部固定下のマウスをオペラント条件づけで訓練し, 報酬確率を操作して消去中を行い, その反応の変化を解析することで, 報酬確率によって消去バーストの有無が制御されていることが明らかになった. さらに同実験でも, 瞳孔サイズの計測を行うことで, 消去の開始に伴って, 一時的な瞳孔サイズの上昇が確認された. 瞳孔サイズは, 学習や予測, あるいは反応の出力に関わる, 感覚運動処理の能動的な調節因子であると考えられる. こうした瞳孔サイズの動態や好奇心駆動型強化学習によって, 環境へと能動的に働きかけることで, 自身の学習を促す動物の新たな一側面を明らかにした. さらに, 強化学習によるモデル化を通して, 学習心理学における理論や, 神経科学との関連を浮き彫りにした.

研究 3 では, 実験とは無関係に生じる他行動を, 強化学習における状態として取り入れることで, 数秒から数十秒の時間スケールで反応を解析することで立ち現れる, バウト・休止パターンのモデル化を行った. ここでは, エージェントの行動をオペラント反応と他行動の選択行動とみなし, さらにその間での切り替えにコストを仮定することによって, バウト・休止パターンをシミュレーション上で再現することに成功した. 強化学習によって, バウト・休止パターンの背後にある機

械論的なメカニズムをモデル化することで, 行動の記述だけに留まらず, バウト・休止パターンの生成を担う神経基盤に関する仮説を提供した.

研究 4 では, 研究 3 で導入した他行動という視点を拡張して, 動物の行動を, 反応が相互結合したネットワークという, マクロなスケールで捉えることを提案し, 習慣形成という現象を, ネットワークの構造変化によって説明した. 従来では, 習慣形成を, 目的志向システムから習慣システムへの移行という, 単一の反応に対する 2 つの独立したシステムの競合的な支配として考えられていたが (Daw et al., 2005; Perez and Dickinson, 2020), ネットワークというマクロな視点を導入することによって, これらのシステム間の相互作用を階層的に捉えることができることを示した. さらに, 提案モデルから, 既存の理論と, 相互排他的な実験事態を予測し, 習慣形成を促す心理学的なメカニズムを明らかにする道筋を切り拓いた. さらに, 提案モデルによってもたらされる行動のマクロな視点は, 現代の行動計測技術の発展によってもたらされる, 大規模な行動データを扱う上で, 学習心理学における計測手法の発展を促す一助となることが期待される.

総括

強化学習による分野間の相互乗り入れ

行動分析学と学習心理学は, その哲学的な背景故に対立的に捉えられることもあったが (丹野, 2019), 行動分析学においても内部状態の導入を肯定する意見があり (Staddon, 2014), 学習心理学においても, 環境と行動の数理的な記述によって対立軸を解消しうることが主張されてきた (澤, 2021). 本研究では, 強化学習という枠組みの下で, 行動分析学と学習心理学のそれぞれで扱われてきた現象を, 包括的に扱った. このように, 強化学習を通じて, 行動分析学と学習心理学は相互に協力しながら発展することが可能となる.

行動分析学と学習心理学だけに限らず, 強化学習や数理的な行動の記述は, 用語 (term) によって隔てられた他の領域との架け橋となりうる. 例えば, 研究 4 で議論

したように, 行動ネットワーク上でのエージェントの遷移は, プランニングとの類似性が認められ, ネットワークの構造変化によって生じる習慣形成は, プランニングに要する計算コストを削減していた. これは, アルゴリズムの上で, 作業記憶への負荷が習慣形成に及ぼす影響を検討した研究や, 認知的な計算リソースの削減として習慣形成をみなすこととの, 親和性が認められる. さらに, 代表的な強化学習のアルゴリズムである Q-learning では, 行動価値関数が逐次更新されるため, 過去の経験を明示的に保存する必要はなく, 学習率 α によって, 直前の行動価値関数と直前の経験の重みが決定される. 最も極端な例では, $\alpha = 0$ では, エージェントは直近の結果によってのみ行動を決定するようになる. 言い換えれば, 過去の極端に記憶のストレージが小さいエージェントと解釈することができる. この解釈は決して表面的なものではなく, Q-learning は行動価値関数の逐次更新とみなせる一方で, 単純な式変形によって, 過去の報酬を指数関数的に重みづけたものの和として行動価値関数を表すこともできる. ここでは学習率 α によって指数関数の減衰速度が決定されるため, 学習率をある種の記憶のストレージとして解釈することも可能だろう. このように, 数理的な記述は, 具体的な用語を抽象化することで, 異なる文脈での解釈を促すことができ, それによって異なる視点から現象を捉えたり, あるいは一見すると異なる現象の間に共通性を見出したりすることにつながる. 既に心理学で応用がされつつある強化学習には, 実験心理学の異なる分野を繋げる役割が期待される.

さらに, 強化学習という枠組みは神経科学でも採用されており, アルゴリズムの神経実装や (Schultz et al., 1997; Dabney et al., 2021), パラメータと神経修飾物質の関係 (Schweighofer and Doya, 2003; Wang et al., 2016) が明らかになっている. 強化学習を採用するメリットの 1 つは, この神経科学との接続である. 個々の研究で, その詳細について議論したように, 強化学習のような機械論的なモデルを行動の背後に想定することで, 行動レベルの研究から, その背後にある神経基盤についての仮説を提供することができる. 少なくとも現在は, 行動科学から神経科学への仮説や, 現象の提供が主となるかもしれないが, 神経活動の操作や神経活動からの知覚体験やイメージのデコーディングといった技術は, 今後の行動科学の発

展に寄与しうる可能性を秘めている. こうした背景を踏まえると, 強化学習のように分野を超えて応用が進む枠組みを採用することで, 行動分析学や学習心理学は異分野との相互作用によって, より拡張, 発展することが期待される.

計測技術に期待されるもの

本博士論文では, 強化学習という理論的枠組みを学習心理学と行動分析学で扱われてきた現象へと適用するだけでなく, 計測という側面でも計算論的手法を採用した. こうした計測技術の発展は行動分析学と学習心理学に何をもたらすだろうか. まずは研究 1 と 2 で計測, 解析した瞳孔を例に考えてみる. 瞳孔サイズの変化は, 覚醒度, 注意, ワーキングメモリ, 社会的警戒, 選択課題における選択肢の価値, 不確実性など, 様々な内的状態と関連している (Ebitz et al., 2014; Ebitz and Platt, 2015; Finke et al., 2021; Joshi and Gold, 2020; Larsen and Waters, 2018; Van Slooten et al., 2018; Vincent et al., 2019; Zénon, 2019). 学習の文脈では, 瞳孔の大きさと, TD 学習における予測誤差 (Sutton and Barto, 2018)や, Pearce-Hall モデル (Pearce and Hall, 1980) における刺激への注意といった学習理論との関係も議論されている (Koenig et al., 2017; Pietrock et al., 2019; Vincent et al., 2019). こうした事実を踏まえると, 瞳孔サイズの変化は単に刺激に対する応答を反映したものではなく, 外部からの刺激を受けて, 自身の与えられる感覚入力の変調など, より能動的な過程を反映していると考えられる. 研究 1 では, 報酬予測によって瞳孔サイズが拡大することを明らかにしたが, ここには報酬予測的な刺激への注意や, 報酬予測に基づいて覚醒度を変調することで, 学習や報酬の処理を促すような能動的なプロセスであると解釈することもできる. 研究 2 では, 消去の開始に伴って, 消去バーストと瞳孔サイズの上昇が観察された. 好奇心駆動型強化学習という枠組みでは, 消去バーストは, 急激な予測誤差の上昇を検出することで, 一時的な反応の増加を促すことで説明される. あえて口語的に表現するならば, 報酬が出なくなったことを「確かめる」反応として消去バーストを説明する. 従って, 研究 2 でも瞳孔がある種の好奇心のような, 動物が環境へと能動的に関わるプロセスを

反映したものとして解釈することができる. これらの研究では, 瞳孔という指標を通して, 動物が能動的に環境へと働きかける, 学習の新たな側面が描き出された.

研究 3 と 4 は, シミュレーション実験のみで構成されており, 一見すると計測とは無縁に思えるかもしれないが, これら 2 つの研究に共通する他行動というアイデアは, 行動計測の問題と密接な関わりを持つ. 行動分析学や学習心理学では, 実験的に定義された反応のみが計測, 解析されてきた歴史があり, こうした反応の性質, 例えば反応率や消去抵抗, と環境上の変数の関係が主として注目されてきた. その一方で, 近年の網羅的な行動解析技術の登場によって, 実験的に定義されない, しかし, 現実の生物が従事する多くの反応が計測できるようになった. こうした反応は, 従来の学習心理学や行動分析学において, 扱う必要のない対象なのだろうか. 少なくとも本能的逸脱 (Breland and Breland, 1961) やスケジュール誘導性行動 (Falk, 1966; Gentry, 1968; Hymowitz, 1971; Levitsky, 1968) といった現象が示すように, 動物が生来より有する, 一見すると実験とは無関係な行動は, 学習性の行動と全く無関係のものではない. そこで, 研究 3 と 4 では, この他行動を明示的に取り扱うことで, バウト・休止パターンや習慣形成といった, 行動現象を従来とは異なる視点で捉えることに成功した. 実際に, これらの現象において, 他行動の存在がどこまで重要であるかは現在のところは明らかになっておらず, 今後の研究によって明らかになることが期待される. しかし, 研究 3 と 4 は, 計測技術の発展によってもたらされる, 大規模な行動データを目の前に, どのように行動の捉えることができるか, その指針の 1 つを示すものである.

このように計測技術の発展によって, 瞳孔のような動物の微細な反応から, あらゆる動作の分類による遷移確率構造というマクロな行動指標までが計測可能になっている. そして, それらの指標によって, 動物が能動的に環境に働きかける側面や, 実験だけでなく, 多様な反応に従事する動物本来の姿など, 動物の様々な側面を捉えることができるようになる. 計測における計算論的手法の活用によって, 新たな指標の計測を通して, これまでの指標では捉えられなかった, 動物の側面を捉えることができるようになる.

学習心理学と行動分析学における計算論的手法の意義

本博士論文では, 強化学習という理論的枠組みによって, 様々な時間スケールでパブロフ型条件づけとオペラント条件づけが関わる行動現象を捉えることを示し, 行動を記述する枠組みとしての包括性を示した. また, 機械学習による, 計測技術の発展によって, 動物の行動を異なる側面で捉えることができることを示した. 動物は環境からの入力を一方的に受けて, 行動を変容するだけの風見鶏ではなく, 入力から出力, そしてその入出力を繋ぐ内的な過程が混然一体となったシステムである. 計算論的手法は, 分野や現象を捉えるスケールを穿ち, そして動物の行動を, あらゆる角度, 解像度で可視化することで, 全体としての動物の行動を捉えるための強力なツールとなることが期待される.

参考文献

Adams, C. D., & Dickinson, A. (1981). Instrumental responding following reinforcer devaluation. *The Quarterly Journal of Experimental Psychology Section B*, *33*(2b), 109-121.

Ainslie, G. W. (1974). Impulse control in pigeons 1. *Journal of the experimental analysis of behavior*, *21*(3), 485-489.

Aldridge, J. W., & Berridge, K. C. (1998). Coding of serial order by neostriatal neurons: a “natural action” approach to movement sequence. *Journal of Neuroscience*, *18*(7), 2777-2787.

Arruda, M. D. O. V., Soares, P. M., Honório, J. E. R., Lima, R. C. D. S., Chaves, E. M. C., Lobato, R. D. F. G., ... & Vasconcelos, S. M. M. (2008). Activities of the antipsychotic drugs haloperidol and risperidone on behavioural effects induced by ketamine in mice. *Scientia Pharmaceutica, 76*(4), 673-688.

Azrin, N. H., Hutchinson, R. R., & Hake, D. F. (1966). EXTINCTION-INDUCED AGGRESSION[1]. *Journal of the Experimental Analysis of Behavior*, *9*(3), 191-204.

Anderson, D. J., & Perona, P. (2014). Toward a science of computational ethology. *Neuron*, *84*(1), 18-31.

Bannai, M., Fish, E. W., Faccidomo, S., & Miczek, K. A. (2007). Anti-aggressive effects of agonists at 5-HT1B receptors in the dorsal raphe nucleus of mice. *Psychopharmacology*, *193*(2), 295-304.

Barabasi, A. L. (2005). The origin of bursts and heavy tails in human dynamics. *Nature*, *435*(7039), 207-211.

Barabasi, A. L. (2016) *Network Science.* Cambridge University Press.

Barter, J. W., Li, S., Lu, D., Bartholomew, R. A., Rossi, M. A., Shoemaker, C. T., ... & Yin, H. H. (2015). Beyond reward prediction errors: the role of dopamine in movement kinematics. *Frontiers in Integrative Neuroscience*, *9*, 39.

Baum, W. M. (1973). The correlation-based law of effect[1]. *Journal of the Experimental Analysis of Behavior*, *20*(1), 137-153.

Baum, W. M. (1981). Optimization and the matching law as accounts of instrumental behavior. *Journal of the Experimental Analysis of Behavior*, *36*(3), 387-403.

Baum, W. M. (1993). Performances on ratio and interval schedules of reinforcement: Data and theory. *Journal of the Experimental Analysis of Behavior, 59*(2), 245-264.

Baum, W. M., & Grace, R. C. (2020). Matching theory and induction explain operant performance. *Journal of the Experimental Analysis of Behavior*, *113*(2), 390-418.

Baum, W. M., & Rachlin, H. C. (1969). Choice as time allocation[1]. *Journal of the Experimental Analysis of Behavior*, *12*(6), 861-874.

Balleine, B. W., & Dickinson, A. (1998). Goal-directed instrumental action: contingency and incentive learning and their cortical substrates. *Neuropharmacology*, *37*(4-5), 407-419.

Balleine, B. W., & O'doherty, J. P. (2010). Human and rodent homologies in action control: corticostriatal determinants of goal-directed and habitual action. *Neuropsychopharmacology, 35*(1), 48-69.

Bellemare, M., Srinivasan, S., Ostrovski, G., Schaul, T., Saxton, D., & Munos, R. (2016). Unifying count-based exploration and intrinsic motivation. *Advances in Neural Information Processing Systems*, 29.

Bennett, J. A., Hughes, C. E., & Pitts, R. C. (2007). Effects of methamphetamine on response rate: A microstructural analysis. *Behavioural Processes*, *75*(2), 199-205.

Bernardi, M. M., De Souza, H., & Neto, J. P. (1981). Effects of single and long-term haloperidol administration on open field behavior of rats. *Psychopharmacology*, *73*(2), 171-175.

Bernstein, J. G., & Boyden, E. S. (2011). Optogenetic tools for analyzing the neural circuits of behavior. *Trends in Cognitive Sciences*, *15*(12), 592-600.

Bowers, M. T., Hill, J., & Palya, W. L. (2008). Interresponse time structures in variable-ratio and variable-interval schedules. *Journal of the Experimental Analysis of Behavior*, *90*(3), 345-362.

Brackney, R. J., Cheung, T. H., Neisewander, J. L., & Sanabria, F. (2011). The isolation of motivational, motoric, and schedule effects on operant performance: a modeling approach. *Journal of the Experimental Analysis of Behavior*, *96*(1), 17-38.

Brackney, R. J., Cheung, T. H., Herbst, K., Hill, J. C., & Sanabria, F. (2012). Extinction learning deficit in a rodent model of attention-deficit hyperactivity disorder. *Behavioral and Brain Functions*, *8*(1), 1-10.

Breland, K., & Breland, M. (1961). The misbehavior of organisms. *American Psychologist*, *16*(11), 681.

Brown, P. L., & Jenkins, H. M. (1968). AUTO-SHAPING OF THE PIGEON'S KEY-PECK 1. *Journal of the Experimental Analysis of Behavior*, *11*(1), 1-8.

Carmena, J. M., Lebedev, M. A., Crist, R. E., O'Doherty, J. E., Santucci, D. M., Dimitrov, D. F., ... & Nicolelis, M. A. L. (2003). Learning to control a brain–machine interface for reaching and grasping by primates. *PLoS Biology*, *1*(2), e42.

Carpenter, B., Gelman, A., Hoffman, M. D., Lee, D., Goodrich, B., Betancourt, M., ... & Riddell, A. (2017). Stan: A probabilistic programming language. *Journal of Statistical Software*, *76*(1).

Chang, A. J., Wang, L., Lucantonio, F., Adams, M., Lemire, A. L., Dudman, J. T., & Cohen, J. Y. (2021). Neuron-type specificity of dorsal raphe projections to ventral tegmental area. *BioRxiv*.

Chapin, J. K., Moxon, K. A., Markowitz, R. S., & Nicolelis, M. A. (1999). Real-time control of a robot arm using simultaneously recorded neurons in the motor cortex. *Nature Neuroscience*, *2*(7), 664-670.

Cohen, J. Y., Haesler, S., Vong, L., Lowell, B. B., & Uchida, N. (2012). Neuron-type-specific signals for reward and punishment in the ventral tegmental area. *Nature*, *482*(7383), 85-88.

Colwill, R. M., & Rescorla, R. A. (1985). Instrumental responding remains sensitive to reinforcer devaluation after extensive training. *Journal of Experimental Psychology: Animal Behavior Processes*, *11*(4), 520.

Colwill, R. M., & Rescorla, R. A. (1988). Associations between the discriminative stimulus and the reinforcer in instrumental learning. *Journal of Experimental Psychology: Animal Behavior Processes*, *14*(2), 155.

Conceição, I. M., & Frussa-Filho, R. (1996). Effects of microgram doses of haloperidol on open-field behavior in mice. *Pharmacology Biochemistry and Behavior*, *53*(4), 833-838.

Corbit, L. H., Chieng, B. C., & Balleine, B. W. (2014). Effects of repeated cocaine exposure on habit learning and reversal by N-acetylcysteine. *Neuropsychopharmacology*, *39*(8), 1893-1901.

Dabney, W., Kurth-Nelson, Z., Uchida, N., Starkweather, C. K., Hassabis, D., Munos, R., & Botvinick, M. (2020). A distributional code for value in dopamine-based reinforcement learning. *Nature*, *577*(7792), 671-675.

Datta, S. R., Anderson, D. J., Branson, K., Perona, P., & Leifer, A. (2019). Computational neuroethology: a call to action. *Neuron*, *104*(1), 11-24.

Daw, N. D., Niv, Y., & Dayan, P. (2005). Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioral control. *Nature Neuroscience*, *8*(12), 1704-1711.

Daw, N. D., Gershman, S. J., Seymour, B., Dayan, P., & Dolan, R. J. (2011). Model-based influences on humans' choices and striatal prediction errors. *Neuron, 69*(6), 1204-1215.

Dayan, P., & Daw, N. D. (2008). Decision theory, reinforcement learning, and the brain. *Cognitive, Affective, & Behavioral Neuroscience*, *8* (4), 429-453.

Decker, J. H., Otto, A. R., Daw, N. D., & Hartley, C. A. (2016). From creatures of habit to goal-directed learners: Tracking the developmental emergence of model-based reinforcement learning. *Psychological Science*, *27*(6), 848-858.

De Houwer, J. (2019). On how definitions of habits can complicate habit research. *Frontiers in Psychology*, *10*, 2642.

DeRusso, A., Fan, D., Gupta, J., Shelest, O., Costa, R. M., & Yin, H. H. (2010). Instrumental uncertainty as a determinant of behavior under interval schedules of reinforcement. *Frontiers in Integrative Neuroscience*, *4*, 17.

Dezfouli, A., & Balleine, B. W. (2012). Habits, action sequences and reinforcement learning. *European Journal of Neuroscience*, *35*(7), 1036-1051.

Dezfouli, A., & Balleine, B. W. (2013). Actions, action sequences and habits: evidence that goal-directed and habitual action control are hierarchically organized. *PLoS Computational Biology*, *9*(12), e1003364.

Dezfouli, A., Lingawi, N. W., & Balleine, B. W. (2014). Habits as action sequences: hierarchical action control and changes in outcome value. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *369*(1655), 20130482.

Dickinson, A. (1985). Actions and habits: the development of behavioural autonomy. *Philosophical Transactions of the Royal Society of London. B, Biological Sciences*, *308*(1135), 67-78.

Dickinson, A., Nicholas, D. J., & Adams, C. D. (1983). The effect of the instrumental training contingency on susceptibility to reinforcer devaluation. *The Quarterly Journal of Experimental Psychology*, *35*(1), 35-51.

Dijkstra, E. W. (1959). *Communication with an automatic computer* (Doctoral dissertation, Excelsior).

Doya, K. (2002). Metalearning and neuromodulation. *Neural Networks*, *15*(4-6), 495-506.

Ebitz, R. B. & Platt, M. L. (2015). Neuronal activity in primate dorsal anterior cingulate cortex signals task conflict and predicts adjustments in pupil-linked arousal. *Neuron*, *85*(3), 628-40.

Ebitz, R. B. & Moore, T. (2019). Both a Gauge and a Filter: Cognitive Modulations of Pupil Size. *Frontiers in Neurology*, *9*, 1190.

Ebitz, R. B., Pearson, J. M., & Platt, M. L. (2014). Pupil size and social vigilance in rhesus macaques. *Frontiers in Neuroscience*, *8*, 100.

Esber, G. R., & Haselgrove, M. (2011). Reconciling the influence of predictiveness and uncertainty on stimulus salience: a model of attention in associative learning. *Proceedings of the Royal Society B: Biological Sciences*, *278*(1718), 2553-2561.

Esteves, F., Parra, C., Dimberg, U., & Öhman, A. (1994). Nonconscious associative learning: Pavlovian conditioning of skin conductance responses to masked fear‐relevant facial stimuli. *Psychophysiology, 31*(4), 375-385.

Falk, J. L. (1966). Schedule-induced polydipsia as a function of fixed interval length 1. *Journal of the Experimental Analysis of Behavior*, *9*(1), 37-39.

Fantino, E., Preston, R. A., & Dunn, R. (1993). Delay reduction: Current status. *Journal of the Experimental Analysis of Behavior, 60*(1), 159-169.

Ferster, C. B., & Skinner, B. F. (1957). *Schedules of reinforcement*. Appleton-Century-Crofts.

Finke, J. B., Roesmann, K., Stalder, T., & Klucken, T. (2021). Pupil dilation as an index of Pavlovian conditioning. A systematic review and meta-analysis. *Neuroscience & Biobehavioral Reviews, 130*, 351-368.

Fleshler, M., & Hoffman, H. S. (1962). A progression for generating variable-interval schedules. *Journal of the Experimental Analysis of Behavior*, *5*(4), 529.

Garcia, J., Ervin, F. R., & Koelling, R. A. (1966). Learning with prolonged delay of reinforcement. *Psychonomic Science*, *5*(3), 121-122.

Garr, E., & Delamater, A. R. (2019). Exploring the relationship between actions, habits, and automaticity in an action sequence task. *Learning & Memory, 26*(4), 128-132.

Garr, E., Bushra, B., Tu, N., & Delamater, A. R. (2020). Goal-directed control on interval schedules does not depend on the action–outcome correlation. *Journal of Experimental Psychology: Animal Learning and Cognition, 46*(1), 47.

Ge, H., Xu, K., & Ghahramani, Z. (2018, March). Turing: a language for flexible probabilistic inference. *Proceedings of the Twenty-First International Conference on Artificial Intelligence and Statistics*, in *Proceedings of Machine Learning Research*, *84*,1682-1690.

Gentry, W. D. (1968). FIXED-RATIO SCHEDULE-INDUCED AGGRESSION 1. *Journal of the Experimental Analysis of Behavior*, *11*(6), 813-817.

Gershman, S. J., & Daw, N. D. (2017). Reinforcement learning and episodic memory in humans and animals: an integrative framework. *Annual Review of Psychology*, *68*, 101.

Gilbert, T. F. (1958). Fundamental dimensional properties of the operant. *Psychological Review*, *65*(5), 272.

Gibbon, J., Church, R. M., & Meck, W. H. (1984). Scalar timing in memory. *Annals of the New York Academy of Sciences*, *423*(1), 52-77.

Gläscher, J., Daw, N., Dayan, P., & O'Doherty, J. P. (2010). States versus rewards: dissociable neural prediction error signals underlying model-based and model-free reinforcement learning. *Neuron*, *66*(4), 585-595.

Glimcher, P. W. (2011). Understanding dopamine and reinforcement learning: the dopamine reward prediction error hypothesis. *Proceedings of the National Academy of Sciences*, *108*, 15647-15654.

Graybiel, A. M. (1998). The basal ganglia and chunking of action repertoires. *Neurobiology of Learning and Memory*, *70*(1-2), 119-136.

Graybiel, A. M. (2008). Habits, rituals, and the evaluative brain. *Annual Review of Neuroscience*, *31*, 359-387.

Grenhoff, J., Nisell, M., Ferre, S., Aston-Jones, G., & Svensson, T. H. (1993). Noradrenergic modulation of midbrain dopamine cell firing elicited by stimulation of the locus coeruleus in the rat. *Journal of Neural Transmission/General Section JNT*, *93*(1), 11-25.

Grossman, C. D., Bari, B. A., & Cohen, J. Y. (2022). Serotonin neurons modulate learning rate through uncertainty. *Current Biology, 32*(3), 586-599.

Guthrie, E. R. (1930). Conditioning as a principle of learning. *Psychological Review*, *37*(5), 412.

Guthrie, E. R., & Horton, G. P. (1946). *Cats in a puzzle box.* Rinehart.

Ha, D., & Schmidhuber, J. (2018). World models. *arXiv preprint arXiv:1803.10122*.

Hagberg, A., Swart, P., & S Chult, D. (2008). Exploring network structure, dynamics, and function using NetworkX. *Proceedings of the 7th Python in Science Conference,* 11-15.

Hearst, E., & Jenkins, H. M. (1974). *Sign-tracking: The stimulus-reinforcer relation and directed action*. Psychonomic Society.

Herrnstein, R. J. (1961). Relative and absolute strength of response as a function of frequency of reinforcement. *Journal of the Experimental Analysis of Behavior*, *4*(3), 267.

Herrnstein, R. J. (1970). On the law of effect[1]. *Journal of the Experimental Analysis of Behavior*, *13*(2), 243-266.

Hilgard, E. R., & Bower, G. H. (1966). *Theories of learning (3rd ed.)*. Appleton-Century-Crofts.

Holschbach, M. A., Vitale, E. M., & Lonstein, J. S. (2018). Serotonin-specific lesions of the dorsal raphe disrupt maternal aggression and caregiving in postpartum rats. *Behavioural Brain Research*, *348*, 53-64.

Hong, S., Jhou, T. C., Smith, M., Saleem, K. S., & Hikosaka, O. (2011). Negative reward signals from the lateral habenula to dopamine neurons are mediated by rostromedial tegmental nucleus in primates. *Journal of Neuroscience*, *31*(32), 11457-11471.

Houthooft, R., Chen, X., Duan, Y., Schulman, J., De Turck, F., & Abbeel, P. (2016). Vime: Variational information maximizing exploration. *Advances in Neural Information Processing Systems*, 29.

Hymowitz, N. (1971). Schedule-induced polydipsia and aggression in rats. *Psychonomic Science*, *23*(3), 226-228.

Ikeda, T., & Hikosaka, O. (2003). Reward-dependent gain and bias of visual responses in primate superior colliculus. *Neuron*, *39*(4), 693-700.

Ikeda, T., & Hikosaka, O. (2007). Positive and negative modulation of motor response in primate superior colliculus by reward expectation. *Journal of Neurophysiology*, *98*(6), 3163-3170.

Jenkins, H. M., & Moore, B. R. (1973). THE FORM OF THE AUTO-SHAPED RESPONSE WITH FOOD OR WATER REINFORCERS 1. *Journal of the Experimental Analysis of Behavior*, *20*(2), 163-181.

Joshi, S., & Gold, J. I. (2020). Pupil size as a window on neural substrates of cognition. *Trends in Cognitive Sciences*, *24*(6), 466-480.

Joshi, S., Li, Y., Kalwani, R. M., & Gold, J. I. (2016). Relationships between pupil diameter and neuronal activity in the locus coeruleus, colliculi, and cingulate cortex. *Neuron*, *89*(1), 221-234.

Jurado-Parras, M. T., Safaie, M., Sarno, S., Louis, J., Karoutchi, C., Berret, B., & Robbe, D. (2020). The dorsal striatum energizes motor routines. *Current Biology*, *30*(22), 4362-4372.

Kamitani, Y., & Tong, F. (2005). Decoding the visual and subjective contents of the human brain. *Nature Neuroscience*, *8*(5), 679-685.

Kaneko, S., Niki, Y., Yamada, K., Nasukawa, D., Ujihara, Y., and Toda, K. (2022). Systemic injection of nicotinic acetylcholine receptor antagonist mecamylamine affects licking, eyelid size, and locomotor and autonomic activities but not temporal prediction in male mice. *Molecular Brain, 17*, 77.

Koenig, S., Uengoer, M., & Lachnit, H. (2018). Pupil dilation indicates the coding of past prediction errors: Evidence for attentional learning theory. *Psychophysiology*, *55*(4), e13020.

Katz, B. R., & Lattal, K. A. (2020). What is an extinction burst?: A case study in the analysis of transitional behavior. *Journal of the Experimental Analysis of Behavior*, *115*(1), 129-140.

Kaye, H., & Pearce, J. M. (1987). Hippocampal lesions attenuate latent inhibition and the decline of the orienting response in rats. *The Quarterly Journal of Experimental Psychology Section B*, *39*(2b), 107-125.

Keller, F. S., & Schoenfeld, W. N. (1950). *Principles of psychology: A systematic text in the science of behavior*. Appleton-Century-Crofts

Keramati, M., Smittenaar, P., Dolan, R. J., & Dayan, P. (2016). Adaptive integration of habits into depth-limited planning defines a habitual-goal–directed spectrum. *Proceedings of the National Academy of Sciences, 113*(45), 12868-12873.

Killeen, P. (1975). On the temporal control of behavior. *Psychological Review*, *82*(2), 89.

Killeen, P. R., Hall, S. S., Reilly, M. P., & Kettle, L. C. (2002). Molecular analyses of the principal components of response strength. *Journal of the Experimental Analysis of Behavior*, *78*(2), 127-160.

Killeen, P. R., & Fetterman, J. G. (1988). A behavioral theory of timing. *Psychological Review*, *95*(2), 274.

Korczyn, A. D., & Keren, O. (1980). The effect of dopamine on the pupillary diameter in mice. *Life Sciences*, *26*(10), 757-763.

Kosaki, Y., & Dickinson, A. (2010). Choice and contingency in the development of behavioral autonomy during instrumental conditioning. *Journal of Experimental Psychology: Animal Behavior Processes*, *36*(3), 334.

Kruglanski, A. W., & Szumowska, E. (2020). Habitual behavior is goal-driven. *Perspectives on Psychological Science*, *15*(5), 1256-1271.

Kulubekova, S., & McDowell, J. J. (2008). A computational model of selection by consequences: Log survivor plots. *Behavioural Processes*, *78*(2), 291-296.

Larsen, R. S., & Waters, J. (2018). Neuromodulatory correlates of pupil dilation. *Frontiers in Neural Circuits*, *12*, 21.

Lashley, K. S., & McCarthy, D. A. (1926). The survival of the maze habit after cerebellar injuries. *Journal of Comparative Psychology*, *6*(6), 423.

Lebedev, M. A., Carmena, J. M., O'Doherty, J. E., Zacksenhouse, M., Henriquez, C. S., Principe, J. C., & Nicolelis, M. A. (2005). Cortical ensemble adaptation to represent velocity of an artificial actuator controlled by a brain-machine interface. *Journal of Neuroscience*, *25*(19), 4681-4693.

Lee, C. R., & Margolis, D. J. (2016). Pupil dynamics reflect behavioral choice and learning in a go/nogo tactile decision-making task in mice. *Frontiers in Behavioral Neuroscience*, *10*, 200.

Lerman, D. C., & Iwata, B. A. (1995). Prevalence of the extinction burst and its attenuation during treatment. *Journal of Applied Behavior Analysis*, *28*(1), 93-94.

Lerman, D. C., & Iwata, B. A. (1996). Developing a technology for the use of operant extinction in clinical settings: An examination of basic and applied research. *Journal of Applied Behavior Analysis*, *29*(3), 345-382.

Leon, A., Hernandez, V., Lopez, J., Guzman, I., Quintero, V., Toledo, P., ... & Escamilla, E. (2021). Beyond single discrete responses: An integrative and multidimensional analysis of behavioral dynamics assisted by Machine Learning. *Frontiers in Behavioral Neuroscience*, *15*.

Leuchs, L., Schneider, M., Czisch, M., & Spoormaker, V. I. (2017). Neural correlates of pupil dilation during human fear learning. *Neuroimage*, *147*, 186-197.

Levitsky, D., & Collier, G. (1968). Schedule-induced wheel running. *Physiology & Behavior*, *3*(4), 571-573.

Liao, R. M., & Ko, M. C. (1995). Chronic effects of haloperidol and SCH23390 on operant and licking behaviors in the rat. *Chinese Journal of Physiology*, *38*, 65-74.

Liu, Y., Rodenkirch, C., Moskowitz, N., Schriver, B., & Wang, Q. (2017). Dynamic lateralization of pupil dilation evoked by locus coeruleus activation results from sympathetic, not parasympathetic, contributions. *Cell Reports*, *20*(13), 3099-3112.

Lonsdorf, T. B., Menz, M. M., Andreatta, M., Fullana, M. A., Golkar, A., Haaker, J., ... & Merz, C. J. (2017). Don't fear 'fear conditioning': Methodological considerations for the design and analysis of studies on human fear acquisition, extinction, and return of fear. *Neuroscience & Biobehavioral Reviews, 77*, 247-285.

Lowet, A. S., Zheng, Q., Matias, S., Drugowitsch, J., & Uchida, N. (2020). Distributional reinforcement learning in the brain. *Trends in Neurosciences, 43*(12), 980-997.

Lubow, R. E., & Moore, A. U. (1959). Latent inhibition: the effect of nonreinforced pre-exposure to the conditional stimulus. *Journal of Comparative and Physiological Psychology*, *52*(4), 415.

Machado, A. (1997). Learning the temporal dynamics of behavior. *Psychological Review*, *104*(2), 241.

Mackintosh, N. J. (1975). A theory of attention: Variations in the associability of stimuli with reinforcement. *Psychological Review*, *82*(4), 276.

Markowitz, J. E., Gillis, W. F., Beron, C. C., Neufeld, S. Q., Robertson, K., Bhagat, N. D., ... & Datta, S. R. (2018). The striatum organizes 3D behavior via moment-to-moment action selection. *Cell*, *174*(1), 44-58.

Mathis, A., Mamidanna, P., Cury, K. M., Abe, T., Murthy, V. N., Mathis, M. W., & Bethge, M. (2018). DeepLabCut: markerless pose estimation of user-defined body parts with deep learning. *Nature Neuroscience*, *21*(9), 1281-1289.

Matsui, H., Yamada, K., Sakagami, T., & Tanno, T. (2018). Modeling bout–pause response patterns in variable-ratio and variable-interval schedules using hierarchical Bayesian methodology. *Behavioural Processes*, *157*, 346-353.

Matsumoto, M., & Hikosaka, O. (2007). Lateral habenula as a source of negative reward signals in dopamine neurons. *Nature*, *447*(7148), 1111-1115.

Mazur, J. E. (2016). *Learning and Behavior: Eighth Edition.* Psychology Press.

McDowell, J. J. (2004). A computational model of selection by consequences. *Journal of the Experimental Analysis of Behavior*, *81*(3), 297-317.

Miller, R. R., Barnet, R. C., & Grahame, N. J. (1995). Assessment of the Rescorla-Wagner model. *Psychological Bulletin, 117*(3), 363.

Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., ... & Hassabis, D. (2015). Human-level control through deep reinforcement learning. *Nature*, *518*(7540), 529-533.

Musk, E. (2019). An integrated brain-machine interface platform with thousands of channels. *Journal of Medical Internet Research*, *21*(10), e16194.

Nambu, A. (2004). A new dynamic model of the cortico-basal ganglia loop. *Progress in Brain Research*, *143*, 461-466.

Nelson, A., & Mooney, R. (2016). The basal forebrain and motor cortex provide convergent yet distinct movement-related inputs to the auditory cortex. *Neuron*, *90*(3), 635-648.

Nelson, A., Schneider, D. M., Takatoh, J., Sakurai, K., Wang, F., & Mooney, R. (2013). A circuit for motor cortical modulation of auditory cortical activity. *Journal of Neuroscience*, *33*(36), 14342-14353.

Nevin, J. A., & Grace, R. C. (2000). Behavioral momentum and the law of effect. *Behavioral and Brain Sciences*, *23*(1), 73-90.

Niki, H., & Watanabe, M. (1979). Prefrontal and cingulate unit activity during timing behavior in the monkey. *Brain Research*, *171*(2), 213-224.

Niv, Y. (2009). Reinforcement learning in the brain. *Journal of Mathematical Psychology*, *53*(3), 139-154.

Nonomura, S., Nishizawa, K., Sakai, Y., Kawaguchi, Y., Kato, S., Uchigashima, M., ... & Kimura, M. (2018). Monitoring and updating of action selection for goal-directed behavior through the striatal direct and indirect pathways. *Neuron*, *99*(6), 1302-1314.

Notterman, J. M., Schoenfeld, W. N., & Bersh, P. J. (1952). Conditioned heart rate response in human beings during experimental anxiety. *Journal of Comparative and Physiological Psychology*, 45(1), 1–8.

O'Hare, J. K., Ade, K. K., Sukharnikova, T., Van Hooser, S. D., Palmeri, M. L., Yin, H. H., & Calakos, N. (2016). Pathway-specific striatal substrates for habitual behavior. *Neuron*, *89*(3), 472-479.

Öhman, A., Fredrikson, M., Hugdahl, K., & Rimmö, P.-A. (1976). The premise of equipotentiality in human classical conditioning: Conditioned electrodermal responses to potentially phobic stimuli. *Journal of Experimental Psychology: General, 105*(4), 313–337.

Ojala, K. E., & Bach, D. R. (2020). Measuring learning in human classical threat conditioning: Translational, cognitive and methodological considerations. *Neuroscience & Biobehavioral Reviews, 114*, 96-112.

Olivier, B. (2004). Serotonin and aggression. *Annals of the New York Academy of Sciences*, *1036*(1), 382-392.

Otto, A. R., Raio, C. M., Chiang, A., Phelps, E. A., & Daw, N. D. (2013). Working-memory capacity protects model-based learning from stress. *Proceedings of the National Academy of Sciences, 110*(52), 20941-20946.

Park, J. W., Bhimani, R. V., & Park, J. (2017). Noradrenergic modulation of dopamine transmission evoked by electrical stimulation of the locus coeruleus in the rat brain. *ACS Chemical Neuroscience*, *8*(9), 1913-1924.

Pathak, D., Agrawal, P., Efros, A.A. & Darrell, T.. (2017). Curiosity-driven Exploration by Self-supervised Prediction. *Proceedings of the 34th International Conference on Machine Learning*, in *Proceedings of Machine Learning Research*, *70*, 2778-2787.

Pavlov, I. P. (1927). *Conditioned reflexes: an investigation of the physiological activity of the cerebral cortex*. Oxford Univ. Press.

Pearce, J. M., & Hall, G. (1980). A model for Pavlovian learning: variations in the effectiveness of conditioned but not of unconditioned stimuli. *Psychological Review*, *87*(6), 532.

Pearce, J. M., Kaye, H., & Hall, G. (1982). Predictive accuracy and stimulus associability: Development of a model for Pavlovian learning. *Quantitative Analyses of Behavior*, *3*, 241-256.

Peele, D. B., Casey, J., & Silberberg, A. (1984). Primacy of interresponse-time reinforcement in accounting for rate differences under variable-ratio and variable-interval schedules. *Journal of Experimental Psychology: Animal Behavior Processes*, *10*(2), 149.

Perez, O. D., & Dickinson, A. (2020). A theory of actions and habits: The interaction of rate correlation and contiguity systems in free-operant behavior. *Psychological Review*, *127*(6), 945.

Petter, E. A., Gershman, S. J., & Meck, W. H. (2018). Integrating models of interval timing and reinforcement learning. *Trends in Cognitive Sciences*, *22*(10), 911-922.

Pietrock, C., Ebrahimi, C., Katthagen, T. M., Koch, S. P., Heinz, A., Rothkirch, M., & Schlagenhauf, F. (2019). Pupil dilation as an implicit measure of appetitive Pavlovian learning. *Psychophysiology, 56*(12), e13463.

Podlesnik, C. A., Jimenez-Gomez, C., Ward, R. D., & Shahan, T. A. (2006). Resistance to change of responding maintained by unsignaled delays to reinforcement: A response-bout analysis. *Journal of the Experimental Analysis of Behavior*, *85*(3), 329-347.

Privitera, M., Ferrari, K. D., von Ziegler, L. M., Sturman, O., Duss, S. N., Floriou-Servou, A., ... & Bohacek, J. (2020). A complete pupillometry toolbox for real-time monitoring of locus coeruleus activity in rodents. *Nature Protocols*, *15*(8), 2301-2320.

Redgrave, P., Coizet, V., Comoli, E., McHaffie, J. G., Leriche, M., Vautrelle, N., ... & Overton, P. (2010). Interactions between the midbrain superior colliculus and the basal ganglia. *Frontiers in Neuroanatomy*, *4*, 132.

Reimer, J., Froudarakis, E., Cadwell, C. R., Yatsenko, D., Denfield, G. H., & Tolias, A. S. (2014). Pupil fluctuations track fast switching of cortical states during quiet wakefulness. *Neuron*, *84*(2), 355-362.

Rescorla, R. A. (1988). Pavlovian conditioning: It's not what you think it is. *American Psychologist*, *43*(3), 151-161.

Rescorla, R A., & Wagner, A. R. (1972). A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. In A. H. Black & W. F. Proktsy (Eds.), *Classical conditioning Ⅱ: Current research and theory,* 64-99. Appleton-Century-Crofts

Rossi, M. A., Li, H. E., Lu, D., Kim, I. H., Bartholomew, R. A., Gaidis, E., Barter, J. W., Kim, N., Cai, M. T., Soderling, S. H., & Yin, H. H. (2016). A GABAergic nigrotectal pathway for coordination of drinking behavior. *Nature Neuroscience*, *19*(5), 742–748.

Russek, E. M., Momennejad, I., Botvinick, M. M., Gershman, S. J., & Daw, N. D. (2017). Predictive representations can link model-based reinforcement learning to model-free mechanisms. *PLoS Computational biology*, *13*(9), e1005768.

Sakai, Y., & Fukai, T. (2008). The actor-critic learning is behind the matching law: matching versus optimal behaviors. *Neural computation, 20*(1), 227-251.

Samejima, K., Ueda, Y., Doya, K., & Kimura, M. (2005). Representation of action-specific reward values in the striatum. *Science*, *310*(5752), 1337-1340.

Schmidhuber, J. (1990). Making the World Differentiable: On Using Self-Supervised Fully Recurrent Neural Networks for Dynamic Reinforcement Learning and Planning in Non-Stationary Environments. Technical Report, TR FKI-126-90, Department of Computer Science, Technical University of Munich.

Schmidhuber, J. (1991). A possibility for implementing curiosity and boredom in model-building neural controllers. In *Proceedings of the international conference on simulation of adaptive behavior: From animals to Animats*, 222-227.

Schultz, W., Apicella, P., Scarnati, E., & Ljungberg, T. (1992). Neuronal activity in monkey ventral striatum related to the expectation of reward. *Journal of Neuroscience*, *12*(12), 4595-4610.

Schultz, W., Dayan, P., & Montague, P. R. (1997). A neural substrate of prediction and reward. *Science*, *275*(5306), 1593-1599.

Schwabe, L., & Wolf, O. T. (2009). Stress prompts habit behavior in humans. *Journal of Neuroscience, 29*(22), 7191-7198.

Schweighofer, N., & Doya, K. (2003). Meta-learning in reinforcement learning. *Neural Networks*, *16*(1), 5-9.

Sekar, R., Rybkin, O., Daniilidis, K., Abbeel, P., Hafner, D. & Pathak, D.. (2020). Planning to Explore via Self-Supervised World Models. *Proceedings of the 37th International Conference on Machine Learning*, in *Proceedings of Machine Learning Research*, *119*, 8583-8592.

Shams, L., Kamitani, Y., & Shimojo, S. (2000). What you see is what you hear. *Nature*, *408*(6814), 788-788.

Shenhav, A., Cohen, J. D., & Botvinick, M. M. (2016). Dorsal anterior cingulate cortex and the value of control. *Nature Neuroscience, 19*(10), 1286-1291.

Shidara, M., & Richmond, B. J. (2002). Anterior cingulate: single neuronal signals related to degree of reward expectancy. *Science*, *296*(5573), 1709-1711.

Shull, R. L., Gaynor, S. T., & Grimes, J. A. (2001). Response rate viewed as engagement bouts: Effects of relative reinforcement and schedule type. *Journal of the Experimental Analysis of Behavior*, *75*(3), 247-274.

Shull, R. L., Gaynor, S. T., & Grimes, J. A. (2002). Response rate viewed as engagement bouts: Resistance to extinction. *Journal of the Experimental Analysis of Behavior*, *77*(3), 211-231.

Shull, R. L. (2004). Bouts of responding on variable-interval schedules: Effects of deprivation level. *Journal of the Experimental Analysis of Behavior*, *81*(2), 155-167.

Shull, R. L., Grimes, J. A., & Bennett, J. A. (2004). Bouts of responding: The relation between bout rate and the rate of variable-interval reinforcement. *Journal of the Experimental Analysis of Behavior*, *81*(1), 65-83.

Silver, D., Huang, A., Maddison, C. J., Guez, A., Sifre, L., Van Den Driessche, G., ... & Hassabis, D. (2016). Mastering the game of Go with deep neural networks and tree search. *Nature*, *529*(7587), 484-489.

Skinner, B. F. (1953). *Science and human behavior.* MacMillan.

Skinner, B. F. (1956). A case history in scientific method. *American psychologist*, *11*(5), 221.

Skinner, B. F. (1981). Selection by consequences. *Science*, 213, 501-504.

Skinner, B. F. (1999). Are theories of learning is necessary?. In B. F. Skinner (Ed.), *Cumulative record (B. F. Skinner Foundation Reprint Series)* (pp. 69-100). Cambridge, MA: B. F. Skinner Foundation. (Original work published 1950 *Psychological Review*, 57, 193-216)

Sparks, D. L. (1986). Translation of sensory signals into commands for control of saccadic eye movements: role of primate superior colliculus. *Physiological Reviews*, *66*(1), 118-171.

Sorribes, A., Armendariz, B. G., Lopez-Pigozzi, D., Murga, C., & de Polavieja, G. G. (2011). The origin of behavioral bursts in decision-making circuitry. *PLoS Computational Biology*, *7*(6), e1002075.

Small, W. S. (1901). Experimental study of the mental processes of the rat. II. *The American Journal of Psychology*, 206-239.

Smith, T. T., McLean, A. P., Shull, R. L., Hughes, C. E., & Pitts, R. C. (2014). Concurrent performance as bouts of behavior. *Journal of the Experimental Analysis of Behavior*, *102*(1), 102-125.

Staddon, J. (2014). *The new behaviorism*. Psychology Press.

Staddon, J. E. R., & Higa, J. J. (1996). Multiple time scales in simple habituation. *Psychological Review*, *103*(4), 720.

Staddon, J. E. R., & Higa, J. J. (1999). Time and memory: Towards a pacemaker-free theory of interval timing. *Journal of the Experimental Analysis of Behavior*, *71*(2), 215-251.

Staddon, J. E., & Simmelhag, V. L. (1971). The "supersitition" experiment: A reexamination of its implications for the principles of adaptive behavior. *Psychological Review, 78*(1), 3–43.

Stern, J. M., & Taylor, L. A. (1991). Haloperidol inhibits maternal retrieval and licking, but enhances nursing behavior and litter weight gains in lactating rats. *Journal of Neuroendocrinology*, *3*(6), 591-596.

Strömbom, U. Antagonism by haloperidol of locomotor depression induced by small doses of apomorphine. *Journal of Neural Transmission 40*, 191–194 (1977).

Sutherland, N. S., & Mackintosh, N. J. (2016). *Mechanisms of animal discrimination learning*. Academic Press.

Sutton, R. S., & Barto, A. G. (2018). *Reinforcement learning: An introduction*. MIT press.

Tang, C., Pawlak, A. P., Prokopenko, V., & West, M. O. (2007). Changes in activity of the striatum during formation of a motor habit. *European Journal of Neuroscience*, *25*(4), 1212-1227.

Tanno, T. (2016). Response-bout analysis of interresponse times in variable-ratio and variable-interval schedules. *Behavioural Processes*, *132*, 12-21.

Tanno, T., & Silberberg, A. (2012). The copyist model of response emission. *Psychonomic Bulletin & Review*, *19*(5), 759-778.

Thorndike, E. L. (1898). Animal intelligence: An experimental study of the associative processes in animals. *The Psychological Review: Monograph Supplements*, *2*(4), 1-109.

Tian, J., Huang, R., Cohen, J. Y., Osakada, F., Kobak, D., Machens, C. K., ... & Watabe-Uchida, M. (2016). Distributed and mixed information in monosynaptic inputs to dopamine neurons. *Neuron*, *91*(6), 1374-1389.

Tolkamp, B. J., & Kyriazakis, I. (1999). To split behaviour into bouts, log-transform the intervals. *Animal Behaviour*, *57*(4), 807-817.

Tolman, E. C., & Honzik, C. H. (1930). Introduction and removal of reward, and maze performance in rats. *University of California publications in psychology.*

Toda, K., Lusk, N. A., Watson, G. D., Kim, N., Lu, D., Li, H. E., ... & Yin, H. H. (2017). Nigrotectal stimulation stops interval timing in mice. *Current Biology*, *27*(24), 3763-3770.

Van Slooten, J. C., Jahfari, S., Knapen, T., & Theeuwes, J. (2018). How pupil responses track value-based decision-making during and after reinforcement learning. *PLoS Computational Biology*, *14*(11), e1006632.

Vincent, P., Parr, T., Benrimoh, D., & Friston, K. J. (2019). With an eye on uncertainty: Modelling pupillary responses to environmental volatility. *PLoS Computational Biology*, *15*(7), e1007126.

Vinyals, O., Babuschkin, I., Czarnecki, W. M., Mathieu, M., Dudzik, A., Chung, J., ... & Silver, D. (2019). Grandmaster level in StarCraft II using multi-agent reinforcement learning. *Nature*, *575*(7782), 350-354.

Vázquez, A., Oliveira, J. G., Dezsö, Z., Goh, K. I., Kondor, I., & Barabási, A. L. (2006). Modeling bursts and heavy tails in human dynamics. *Physical Review E*, *73*(3), 036127.

Wang, J. X., Kurth-Nelson, Z., Tirumala, D., Soyer, H., Leibo, J. Z., Munos, R., ... & Botvinick, M. (2016). Learning to reinforcement learn. *arXiv preprint arXiv:1611.05763*.

Wang, H., Ortega, H. K., Atilgan, H., Murphy, C. E., & Kwan, A. C. (2022). Pupil Correlates of Decision Variables in Mice Playing a Competitive Mixed-Strategy Game. *eNeuro*, 9(2), 0457-21.

Watanabe, M. (1996). Reward expectancy in primate prefrental neurons. *Nature*, *382*(6592), 629-632.

Watanabe, S., & Opper, M. (2010). Asymptotic equivalence of Bayes cross validation and widely applicable information criterion in singular learning theory. *Journal of Machine Learning Research*, *11*(12).

Wearden, J. H., & Clark, R. B. (1988). Interresponse-time reinforcement and behavior under aperiodic reinforcement schedules: A case study using computer modeling. *Journal of Experimental Psychology: Animal Behavior Processes*, *14*(2), 200.

Wise, R. A. (2004). Dopamine, learning and motivation. *Nature Reviews Neuroscience*, *5*(6), 483-494.

Wiltschko, A. B., Tsukahara, T., Zeine, A., Anyoha, R., Gillis, W. F., Markowitz, J. E., ... & Datta, S. R. (2020). Revealing the structure of pharmacobehavioral space through motion sequencing. *Nature Neuroscience*, *23*(11), 1433-1443.

Wood, D. M., & Obrist, P. A. (1964). Effects of controlled and uncontrolled respiration on the conditioned heart rate response in humans. *Journal of Experimental Psychology, 68*(3), 221–229.

Yamada, K., & Kanemura, A. (2020). Simulating bout-and-pause patterns with reinforcement learning. *Plos One*, *15*(11), e0242201.

Yin, H. H., Knowlton, B. J., & Balleine, B. B. (2004). Lesions of dorsolateral striatum preserve outcome expectancy but disrupt habit formation in instrumental learning. *European Journal of Neuroscience*, 19, 181–189.

Yin, H. H., Ostlund, S. B., Knowlton, B. J., & Balleine, B. W. (2005). The role of the dorsomedial striatum in instrumental conditioning. *European Journal of Neuroscience*, *22*(2), 513-523.

Yin, H. H., & Knowlton, B. J. (2006). The role of the basal ganglia in habit formation. *Nature Reviews Neuroscience*, *7*(6), 464-476.

Yin, H. H. (2010). The sensorimotor striatum is necessary for serial order learning. *Journal of Neuroscience*, *30*(44), 14719-14723.

Zénon, A. (2019). Eye pupil signals information gain. *Proceedings of the Royal Society B*, *286*(1911), 20191593.

鮫島和行 (2022). 機械学習. 澤幸祐 (編), *手を動かしながら学ぶ学習心理学*, (pp. 95-118). 朝倉書店.

澤幸祐. (2021). 関数分析を通じた徹底的行動主義と方法論的行動主義の接続. *行動分析学研究, 35*(2), 158-164.

丹野貴行. (2019). 徹底的行動主義について (特集 坂上貴之教授 退職記念号). *哲學, 142*, 9-42.

松井大. (2021). 行動の数量化とモデル化: オペラント行動のバウト―休止に関する時系列モデルを例に―. *基礎心理学研究, 40*(1), 43-49.

関連業績

投稿論文

1. Matsui, H., **Yamada, K**., Sakagami. T., & Tanno. T. (2018). Modeling bout–pause response patterns in variable-ratio and variable-interval schedules using hierarchical Bayesian methodology. *Behavioural Processes,* 157. 査読あり
2. **Yamada, K.** & Kanemura, A. (2020) Simulating bout and pause patterns with reinforcement learning. *PLoS ONE*. 15(11), e0242201. 査読あり
3. **Yamada, K.** & Toda, K. (2021). Habit formation viewed as structural change in the behavioral network. *bioRxiv*. 査読なし
4. Nasukawa, D., **Yamada, K**., Hirakata, H., Tamura, R., Yatagai, S., Ujihara, Y. & Toda, K. (2022). Differential effects of oxytocin receptor antagonist on social rank and other social behavior in mice. *bioRxiv*. 査読なし
5. Kaneko, S., Niki, Y., **Yamada, K.**, Nasukawa, D., Ujihara, Y. & Toda, K. (2022). Systemic injection of nicotinic acetylcholine antagonist mecamylamine affects licking, eyelid size, locomotor and autonomic activities but not temporal prediction in mice. *Molecular Brain, 15*, 77. 査読あり
6. Yamamoto, K., **Yamada, K.**, Ujihara, Y., Yatagai, S. & Toda, K. (2022). Spatio-temporal pavlovian head-fixed reversal learning task for mice. *Molecular Brain, 15*, 78. 査読あり
7. **Yamada, K.** & Toda, K. (2022). Pupillary Dynamics of Mice Performing a Pavlovian Delay Conditioning Task Reflect Reward Predictive Signals. *Frontiers in Systems Neuroscience*. *accepted.* 査読あり

ポスター発表

1. **Yamada, K.**, Matsui, H., & Toda, K. (2021). Curiosity-driven computational model explains extinction bursts. 日本動物心理学会第 81 回大会, 2021 年 11 月.
2. Kaneko, S., Niki, Y., **Yamada, K.**, & Toda, K. (2021). Effects of the nicotinic acetylcholine receptor antagonist on the performance of temporal conditioning in mice. 日本動物心理学会第 81 回大会, 2021 年 11 月.
3. Niki, Y., Ujihara, Y., Yatagai, S., **Yamada, K**. & Toda, K. (2021). Modulation of the pupillary response in mice during the temporal conditioning task. 日本動物心理学会第 81 回大会, 2021 年 11 月.

4. **Yamada, K.** & Toda, K. (2021). Habit formation viewed as structural change in the behavioral network. The 44th Annual Meeting of the Japan Neuroscience Society.
5. **Yamada, K.**, Kanemura, A. & Sakagami, T. (2018). Simulating bout structure based on reinforcement learning 強化学習に基づくバウト構造のシミュレーション. 動物の反応のバウト／休止パターンに関する時系列モデリング. 日本心理学会大会発表論文集 日本心理学会第 82 回大会.
6. Matsui, H., **Yamada, K**., Sakagami. T., & Tanno. T. (2018). Modeling bout-pause response patterns in variable-ratio and variable-interval schedules 動物の反応のバウト／休止パターンに関する時系列モデリング. 日本心理学会大会発表論文集 日本心理学会第 82 回大会.
7. **Yamada, K.**, Ujihara, Y. & Toda, K. Effects of pharmacological manipulations of dopamine receptors on learning and memory in head-fixed mice. 日本動物心理学会第 80 回大会, 鹿児島, 2020 年 11 月.
8. **Yamada, K.** & Toda, K. Understanding the properties of learning by extracting behavioral elements with machine learning in mice. 日本動物心理学会第 79 回大会, 東京, 2019 年 10 月.

謝辞

本博士論文の執筆と実験の実施にあたり, 多くの方々のご指導とご支援を賜りました. ここで皆様に御礼申し上げます.

兎田幸司先生には, 本博士論文の執筆と実験の遂行にあたって, 実質的なご指導を頂きました. 学部で行動分析学と学習心理学を学び, 修士以降では, 兎田先生の下でご指導いただいたことで, 神経科学に関する知見や手技を学びました. それ以上に大きな糧となったのは, 行動分析学と学習心理学が蓄積した膨大な学習と行動に関する知見が, 神経科学においてもいかに重要であるかを学んだことです.

伊澤栄一先生は, 研究上の助言を頂いただけでなく, 折を見て研究以外の事柄についても相談に乗って頂きました. そして, 私の数理的なモデルを通して行動を捉える部分に, 早い段階から気づき, そうした視点からの助言は, 参考になるだけでなく, 私自身非常に考えさせられるものでした. 心理学者として, 数理的に行動を捉える上での考え方を学びました.

梅田聡先生は, ヒト研究と認知神経科学という, 動物の行動を主として扱ってきた私にはない視点から, 数多くの助言を頂きました. 特に大学院の講義を通して, 強化学習を中心とした論文を読む私に対して, 梅田先生から頂いたコメントから, ヒトと動物, そして認知と行動という, 2 つが, 数理的なものの見方を通じて自然と繋がることに気づかされました.

濱口航介先生には, 本博士論文の副査をお引き受け頂き, 多くの助言を頂きました. 特に本博士論文の主題でもある計算論的手法という点について, 頑強なバックグラウンドを持つ濱口先生から頂いたコメントは, 伊澤先生, 梅田先生, そして兎田先生から頂いたものとは, 異なる角度からのものであり, 非常に貴重なものであると同時に心強いものでした.

私の学部時代にご指導いただいた坂上貴之先生にも, ここで感謝の意を表します. 坂上先生の下で行動分析学について学んだことは, 学習心理学や強化学習, 神経科

学を学んだ今でも残っているどころか, ますますと自身の根底には行動分析家としての考え方が染みついていることを実感させられます.

講義や学会での議論を通して, 学部当時の私に学習心理学の面白さを伝えて頂いた, 神前裕先生, 澤幸祐先生にも御礼申し上げます. 先生たちの講義や議論の中で, 当時, 行動分析学しか知らなかった私に, 似たようでどこか違う, 新しい世界を紹介頂いたことは, 今に至るまで, 大きな影響力を持つこととなりました.

修士時代の 2 年間, 強化学習や機械学習の基礎を学ぶ時間や機会を提供頂いた兼村厚範氏にも感謝申し上げます. 最初の論文投稿にあたって, 一通りのプロセスを懇切丁寧にご指導いただいたことは, 研究者として非常に大きな経験となりました. さらに, 論文執筆と機械学習の勉強にかける時間を頂いたことは, 修士時代の私にとって, 研究・生活上で大きな助けなりました.

学部時代の, まだ右も左も分からない私に, 実験の手技からデータの解析, 論文の読み方など, 研究者としての基礎的な教育を頂いた藤巻峻氏にもお礼申し上げます. 2 年という短い間でしたが, 藤巻さんの毎日実験をする姿からは, 基本的な手技のみならず, 研究者としての姿勢を学びました.

学部時代から今に至るまで, 常に研究に関する議論や数多くの助言を頂いた松井大氏にも多大なる感謝を申し上げます. 松井さんには, 学部時代から行動分析学や学習心理学に関わる議論はもちろんのこと, 松井さんご自身の興味に関わる研究をご紹介いただき, 非常に広い世界が広がっていることを教えて頂きました. そして, 強化学習をはじめとする計算論という世界へ足を踏み出すきっかけとなったのは, 松井さんであることは間違いないでしょう. 本博士論文の執筆や, 一部の研究についても, 様々な助言や議論を頂きました.

同期の盛田和孝氏と氏原勇祐氏にも感謝申し上げます. 動物棟で同じ時を過ごし, 切磋琢磨したことは, 在学中の数年間の中でも最も刺激的で楽しい時間でした. 日頃の動物の世話を協力いただいた研究室のメンバーにも感謝申し上げます.

研究の半数を占めるシミュレーションは計算機やプログラミング言語, エディタの存在なくしてありえませんでした. こうしたツールを使用できるのは, プログラミング言語やエディタ, ライブラリの開発, メンテナンスに携わる世界中のエンジニアによる努力の賜物です. 全世界のエンジニアと有志によって形成されたコミュニティ, そして, そこで生まれた多大な蓄積には多大な御礼と深い尊敬を申し上げます.

最後に, 研究者としての道のりを見守り, あらゆる面において支援頂いた家族に最大の感謝を申し上げます.