

予期的かつ漸進的な身体状態制御のための
恒常性強化学習モデルの研究

藤原 真奈

目次

要旨	3
略語表	4
第一章 序論	5
1.1 研究背景	6
1.1.1 動物の意思決定における「報酬」「罰」の定義と量	6
1.1.2 経験依存的な生理状態の予期的調節	7
1.2 恒常性強化学習モデル	8
1.3 先行研究における恒常性強化学習モデルの課題と本研究での改善	9
第二章 モデルⅠ 予期的ホメオスタシス制御（一次元モデル）	11
2.1 行動パラダイムと対象現象	12
2.2 状態ダイナミクスと予期的制御	14
2.3 ドライブ低減と報酬の定義	16
2.4 予期信号・抑制学習・ゲーティング機構	17
2.5 学習則と報酬帰属	18
2.6 シミュレーション手順と評価指標	20
第三章 結果Ⅰ（一次元モデル）	21
3.1 反応的ホメオスタシス動態のみでの挙動	22
3.2 予期的耐性の漸進的獲得	22
3.3 ゲーティングによる消去と迅速な再獲得	23
3.4 既存 HRL モデルとの比較	23
第四章 モデルⅡ 共有メタ信号による予期的制御（多次元拡張モデル）	29
4.1 共有メタ信号・影響重み・優先度	30
4.2 シミュレーション手順と評価指標	31
4.2.1 単純設定：1 つのメタ信号が 2 変数を制御する場合	31
4.2.2 拡張設定：2 つのメタ信号が 3 変数を制御する場合	32
第五章 結果Ⅱ（多次元拡張モデル）	33
5.1 共有メタ信号制御下におけるトレードオフの創発	34
5.2 共有メタ信号制御下における優先度付けの効果	36
5.3 共有制御下における学習の成功条件と失敗条件	42
第六章 考察	46
6.1. 主要な知見と新規性	47
6.2. 結果の解釈	47
6.3. 病理との関連	48
6.4. 先行研究との比較	49
6.5. 限界と今後の課題	49
6.6. 生理学・モデリングへの含意	51
補遺	52
A1 一次元モデル	53
A1.1 漸進的予期的調節モデルの実装（提案モデル）	53
状態ダイナミクス	53
ドライブ関数と報酬定義	54
予期信号・抑制・ゲーティング	54

学習則と報酬帰属	55
更新信号の強化学習としての解釈	56
A1.2 基準 HRL モデルの実装 (KERAMATI & GUTKIN, 2014)	57
A1.3 感度解析	58
A1.3.1 one-at-a-time (OAT) 法.....	58
A1.3.2 報酬の自己帰属・外因帰属についての議論のための補足 (ω のスweep)	61
A2. 多次元拡張モデル	62
優先度付けのパラメータ化とパラメータスweepについての補足	63
先行モデルのパラメータ m, n と提案モデルの a, u の関係	64
文献.....	65
謝辞.....	72

要旨

生体にとって、体温・血圧・血糖などの生理変数がある程度狭い範囲に保つ恒常性は重要である。変動する環境からの外乱を受けても恒常性を維持するためには、外乱を受けた後に元の状態に戻す「反応的フィードバック調節」だけでなく、外乱に備えて働く「予期的なプロアクティブ調節」の2つの機構があるはずである。本研究では、「予期的なプロアクティブ調節」に着目し、その機構を表現する数理モデルを構築した。

構築したモデルは、恒常性強化学習（homeostatic reinforcement learning; HRL）の基本概念に則っている。HRL システムは、身体の理想状態を設定し、理想に近づく変化を報酬、遠のく変化を罰として学習し、より適切な反応ができる頻度を上げていくというものである。しかし従来の枠組みは、強化学習の機械学習的な側面が強く、生命現象を表現するには不十分な点があった。特に、状態変化を予測的な補償反応を強めていき、漸進的に耐性を獲得していく学習過程の表現が、二値的反応に省略されている点を改善することで、システム内のあらゆる調整機構間の相互連関について示唆が得られると考えた。

本研究では、従来の恒常性強化学習が十分に扱えなかった (i) 試行ごとの漸進的な補償学習と (ii) 消去学習（extinction）と再獲得 を同時に説明するため、補償信号を連続量として更新できる連続 HRL モデルを提案する。具体的には、負の結果を正の結果より強く重みづける 非対称強化（損失優位）と、活性化・抑制の二経路からなる2ユニット型の抑制学習、さらに文脈に応じて抑制の効きを切り替える ゲーティング機構 を組み込むことで、学習された反応が「消去で消える」のではなく「抑制で覆い隠され、条件が戻ると素早く再表出する」ことを表現した。

提案モデルを、例としてエタノール誘発性低体温に対する熱調節古典条件づけ（CS 提示後に US で低体温が生じる）へ適用し、CS 誘発の先行的加温、試行を追った耐性の漸増、消去後の迅速な再獲得を再現した。特に、先行補償がない場合は外乱後に設定点へ復帰できても耐性は獲得されず、損失優位の非対称報酬がなければ学習信号が弱く学習が進まない。また、消去後の“速い戻り”には 文脈依存ゲーティングが本質的で、ゲートがないと抑制が解除されにくく再獲得が鈍化する。これらは、連続的な補償量の試行更新、損失優位評価、抑制記憶+ゲーティングという3要素が、漸進学習と消去・再表出を同時に説明する鍵であることを示した。

さらに本研究は、複数の生理変数が少数の共有制御信号（神経・ホルモン様の“メタ信号”）で同時に制御される多次元系へ拡張し、一つを補償すると別の変数に必然的コストが生じるというトレードオフを明示的に扱う。変数ごとの優先度（重み）が不均衡な条件では、共有制御を介して偏りが増幅し、理想状態近傍へ安定化できない“連鎖的な破綻”が生じうることを示し、こうした失敗様式を自律神経調節不全や ME/CFS などの議論と接続する計算論的視点を与える。以上より、提案枠組みは、単一臓器の適応だけでなく、多臓器・多変数の恒常性破綻をシステムとして理解するための基盤を提供する。

略語表

ATP — Adenosine triphosphate

BA — Basal amygdala

BLA — Basolateral amygdala

CeA — Central amygdala

CS/US — Conditioned stimulus / Unconditioned stimulus

FEP — Free Energy Principle

HRL — Homeostatic Reinforcement Learning

ITCs — Intercalated cell clusters

ME/CFS — Myalgic encephalomyelitis / Chronic fatigue syndrome

OAT — One-at-a-time

RL — Reinforcement Learning

VTA — Ventral tegmental area

第一章

序論

1.1 研究背景

1.1.1 動物の意思決定における「報酬」「罰」の定義と量

動物個体の意思決定を数理モデルで表現したモデルの一つに「強化学習モデル」がある。強化学習モデルは、報酬を得られた行動が強化されてゆく学習モデルである。では、その「報酬」は一体何によって価値づけられているのだろうか。すなわち、何をどの程度の報酬とし、何を罰とするのか、またそれはどのように決定されているのだろうか。

一般にヒトを対象にした心理学実験では金銭を、動物を対象にした実験では甘味溶液や、口渇条件に置いた上での水を報酬として与える。他には、脳の報酬領域（例：腹側被蓋野；VTA）への直接的な電気刺激が強化子として用いられる場合もある。逆に、罰金や侵害刺激（痛みを伴う電気ショックなど）は罰として用いられる。これらに共通する見方として、報酬や罰の「価値」は、環境から恣意的に与えられるラベルではなく、生体が自らの存続条件を維持するうえでの寄与として規定されるということができる。このような見解は、古典的には、飢え・渇き・体温などの内的欠乏（drive）を減らす刺激が報酬となるというドライブ低減と恒常性の立場が出発点となり、段階的に形成されてきた（Cannon, 1932; Hull, 1943）。学習心理学は、快・不快という一次的な経験だけでなく、手がかりや文脈に学習された価値（条件づけ）も行動を強力に導くことを示し、価値は経験と予測で更新される量と捉えられることを明確にした（Robert A. Rescorla & Wagner, 1972）。さらに、神経計算論は、ドパミン作動性信号が「期待と結果のズレ（予測誤差）」に応じて価値表象を更新するという対応関係を与え、価値は固定的な“ご褒美”ではなく、将来の成果を見込む期待値として定式化されるようになった（Montague et al., 1996; Schultz et al., 1997; Sutton & Barto, 2018）。

ここで重要なのは、価値が身体状態によって反転・増減するという事実である。例えば塩分や糖は、欠乏時には強い報酬として機能し得る一方、過剰時には罰に近い価値を取り得る。この性質は、快の経験そのもの（“liking”）と、追求動機（“wanting”）が必ずしも一致しないという観点とも整合する（Berridge et al., 2009; Clark & Bernstein, 2006）。加えて、内部状態、ひいては内受容状態（内部状態に関する神経系の推定状態）が価値づけをゲートするという考え方とも親和性が高い（Barrett & Simmons, 2015; Craig, 2002, 2003; Livneh et al., 2017; Seth, 2013）。すなわち、主観的に感じられる快・不快とは別に、脳が身体の内状態をどう認識しているか（内受容）によって、同一刺激の価値づけが変化し得るということである。

さらに、身体状態制御の枠組みとして「アロスタシス」が提唱されている。アロスタシスは、「変化による安定」を掲げ、外的環境や内的需要の変化を予測的に先回りして調節する枠組みとして展開されてきた（McEwen & Stellar, 1993; Ramsay & Woods, 2014; Sterling & Eyer, 1988）。このように、予測的調節を重視する立場では、価値とは単に「現在の偏差を小さくすること」ではなく、将来の破綻リスクを低減しつつ状態を許容範囲に保つという長期的要請と結び付けて理解される。これに関する規範的な枠組みでは、「将来にわたって生存可能域に留まること」への期待寄与として捉える見方も提唱されている（K. Friston, 2013）。すなわち、価値づけは現在の快・不快のみに還元されるのではなく、予期的調節を通じて長期的な維持可能性を確保する方向へ拡張して定義され得る。

総じて、報酬・罰は「外部から定数として与えられる」のではなく、（i）生理的要請、（ii）学習された予測、（iii）不確実性と将来の選好、（iv）多変量のトレードオフによって決まる状態依存・予測依存の量と捉えられる。本研究は、価値をドライブ（セットポイント偏差）の低減として定義する古典的枠組みに立ちながら、その低減が単なる反応的補償に

とどまらず、手がかりに基づく予期的調節（アロスタシスの側面）として学習され得ることを、恒常性強化学習の形で具体化する。

1.1.2 経験依存的な生理状態の予期的調節

報酬や罰を前節のように捉えると、動的に変化する環境下で恒常性を維持する仕組みは、強化学習（reinforcement learning; RL）的に理解できる。ここでいう RL とは、ある行動の結果として「状態が望ましい方向に近づいたか/遠のいたか」を評価し、その評価に基づいて次回以降の行動選択を更新する一般的な学習原理である。恒常性の文脈では、体温・血糖・水分状態などの内部状態が目標範囲から逸脱すると罰となり、目標範囲に近づく変化は報酬となる。この評価が繰り返しの経験を通じて蓄積されることで、動物は「よりうまく戻せる反応」や「そもそも逸脱を起こしにくい反応」を選びやすくなる。

実際、生物の恒常性維持は、外乱が起きた後に逸脱を補正する反応的制御だけでなく、将来の変化を手がかりから見積もって事前に備える予期的制御にも依存する。「先回り」による調節は、外乱が生じてから大きく崩れて戻すよりも、逸脱自体を小さく抑えられる点で有利である。このような予期的調節は、生得的な要素に加えて、経験を通じて「何が外乱の前触れか」を学ぶことで獲得され得る。

この学習過程の代表例として、パブロフ条件づけによる薬物耐性の形成が挙げられる (Farahbakhsh & Siciliano, 2023; Hou et al., 2023; Le et al., 1979; Poulos & Cappell, 1991; Siegel, 2001, 2005)。薬物を反復投与すると、身体は薬効を相殺する方向へ調節し、次第に効果が減弱するため、同等の治療効果を得るには投与量を増やす必要が生じる。このとき重要なのは、耐性が薬物そのもののだけでなく「薬物投与と一緒に繰り返し経験した手がかり（場所・匂い・手順など）」に依存して表れる点である。すなわち、薬物と連合した経験のない別の文脈へ移ると、学習された補償反応が十分に発現せず、同じ投与量でも過剰な効果が生じうる。これは、身体の調節が手がかり依存的に学習され、文脈によって表出が変わることを示唆する。実験室での具体例としては、体温調節における古典的条件づけがある (Mansfield et al., 1983; Mansfield & Cunningham, 1980)。このパラダイムでは、中性刺激（条件刺激：Conditioned Stimulus, CS）を、低体温を引き起こすエタノール注射（無条件刺激：Unconditioned Stimulus, US）と対提示する。対提示を繰り返すと、動物は「CS の後に低体温が来る」ことを学び、実際に体温が下がる前の CS の時点で体温を上昇させる補償反応を示すようになる。その結果、US による低体温が小さくなり、行動としては耐性の獲得として観察される。以上の例は、恒常性応答が、外乱の前触れとなる手がかりに基づいて予期的に調整され得ることを示している。しかしながら、このような経験依存的・予期的調節を支える計算機構、すなわち、どの量が学習され、どの規則で更新され、文脈依存性がどのように組み込まれるかは、これまで十分に解明されていない。

1.2 恒常性強化学習モデル

手がかり依存的な予期的恒常性調節を扱うために、Keramati と Gutkin は強化学習 (reinforcement learning: RL) の枠組みに基づく恒常性の計算論的意思決定モデルを提案した (Keramati & Gutkin, 2014)。彼らの恒常性強化学習 (homeostatic reinforcement learning: HRL) モデルでは、最適内部状態からの偏差を低減する行動が報酬的で強化され、偏差を増大させる行動は罰として回避されるよう学習が進む (図 1)。このモデルは、水分バランス、体温調節、さらにコカイン依存などの病的状態を含む多様な生理領域へ応用されてきた (Keramati et al., 2017)。また Uchida らは、塩味という条件刺激 (CS) が将来のナトリウム摂取を予測するという観点から、塩分欲求を説明するために HRL を拡張した (Uchida et al., 2022)。

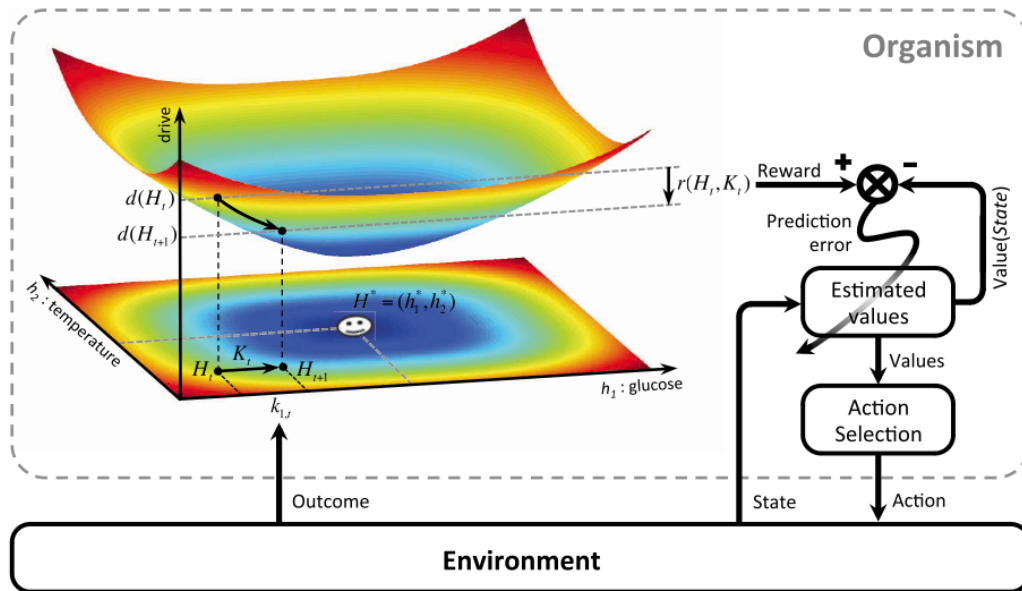


図 1. 二次元の恒常性空間の恒常性強化学習モデル概要 (Keramati & Gutkin, 2014 より転載) . 動物がある行動を行うと、環境から結果として K_t を受け取る。この結果の報酬価値は、その結果が内部状態 H_t を恒常性の設定点 (セットポイント H^*) にどれだけ近づけ、ドライブ水準 (縦軸) をどれだけ低下させるかによって決まる。経験される報酬 $r(H_t, K_t)$ は、その後、強化学習アルゴリズムによって学習される。ここでは、モデルフリー型の強化学習アルゴリズムの例を示しており、実現した報酬と実行した反応に期待される報酬価値を比較することで、報酬予測誤差信号が計算される。この信号は、対応する反応に帰属される主観的価値を更新するために用いられる。さらに、別の選択肢に対する主観的価値が行動選択過程にバイアスを与える。

1.3 先行研究における恒常性強化学習モデルの課題と本研究での改善

原型となる HRL 定式化を生物学的システムへ適用する際には、主に 3 つの限界がある。第一に、従来の HRL は離散的な行動選択を前提としており、補償信号が試行ごとに連続的かつ漸進的に調整されるという生体の特徴を捉えにくい。そこで本研究では、神経回路モデルと HRL 枠組みを統合し、CS によって駆動される補償信号の強度が、試行ごとの強化学習過程を通じて総報酬を最大化するよう更新される構造を導入する。この構造により、内部状態に対する予期的かつ文脈依存的な制御を獲得し、学習済み応答を柔軟に調節できる。

第二に、従来の HRL 枠組みでは、報酬と罰の価値付けが対称であることがしばしば仮定される。しかし、ギャンブルなどの意思決定課題を含む行動・神経科学研究では、損失が同等の利得よりも重く評価される「損失優位 (loss-dominant)」な非対称性が一貫して報告されている (Macpherson et al., 2014; Sokol-Hessner & Rutledge, 2019)。本研究ではこの考え方を内部生理状態の調節へ拡張する。恒常性システムは安全な動作範囲を防衛し、臨界閾値を越える事態を回避するよう組織化されている。その調節は強靱であり、例えば体温は外乱があっても 35–38°C の範囲に厳密に保たれる。しかし制御理論的には、いったん逸脱が生じた後に急速に回復させようとするとオーバーシュートや不安定化のリスクが高まる。臨床的にも、重篤疾患や多臓器不全からの回復はしばしば遷延し、慢性あるいは遷延性の重症疾患において臓器機能障害と長期的後遺障害が持続することが報告されている (Iwashyna, 2012; Voiriot et al., 2022)。したがって本研究では、「閾値越えに対する強い防衛」と「逸脱後の慎重な補正」という制御の非対称性を、損失優位の調節として解釈し、この原理をモデルへ組み込む。

第三に、従来の HRL は、記憶の消去 (extinction) に代表される抑制性学習を十分に扱えない。古典的条件づけにおける消去とは、CS の後に US が提示されなくなったとき、条件反応が徐々に減衰する現象である。しかし消去後であっても、時間経過による自発的回復、US の無信号再提示による再燃、CS-US 対提示再開時の迅速な再獲得など、条件反応の再出現が知られている (Myers & Davis, 2002; Napier et al., 1992; R. A. Rescorla & Heth, 1975)。これらは、消去が元の記憶を単に消去するのではなく、別個の抑制性トレースが形成され、その表出が文脈手がかりによりマスク・解除され得ることを示唆する (Bouton, 2002; Bouton et al., 2006)。本研究ではこの「表出の切替え」を記述するため、文脈依存ゲーティングを導入する。実際、消去表出は神経回路レベルでも、扁桃体介在細胞群 (Intercalated cell clusters, ITCs) を介した扁桃体中心核 (Central amygdala, CeA) 出力の抑制として整理される出力ゲート (Likhtik et al., 2008; Quirk & Mueller, 2008) や、扁桃体基底外側核/扁桃体基底核 (Basolateral amygdala/Basal amygdala, BLA/BA) における細胞集団活動の切替えとして観察される (Herry et al., 2008) ことが報告されており、抑制記憶の形成と表出制御を分けて記述する必要性を支持する。一方、消去を「抑制トレースの形成」として扱うには、活性化と抑制を同時に保持しうる表現が必要であるため、本研究では 2 ユニット構造を採用する。この設計は、哺乳類脳で広く見られる対抗的二経路モチーフ (例: 線条体の直接路/間接路が D1/D2 受容体発現ニューロンを介して行動の促進・抑制に関与する) と整合的であり、連合学習における活性化と抑制の双方を表現するモデル化とも対応する (Li et al., 2016)。これらより、消去と再表出を扱うには、抑制トレースを担う 2 ユニットと、その文脈依存的表出を切り替えるゲーティングの併用が、生物学的に妥当な予期的制御モデルに不可欠である可能性を示している。

以上を踏まえ、本研究ではこれら 3 つの限界を克服する拡張 HRL 枠組みを提案する。提案モデルは、(1) 内的制御信号を試行ごとに漸進的に更新できること、(2) 罰をより重視する非対称な報酬重み付け、(3) 期待された結果が得られない場合に条件反応を抑制するゲーティング機構として実装された抑制性学習を備える。さらに本モデルを多次元の恒常性システ

ムへ拡張し、系全体としての最適化が、局所的な内部状態間のトレードオフの創発や、さらには不適応・病的状態へ至り得ることを示す。

第二章

モデルⅠ 予期的ホメオスタシス制御 (一次元モデル)

自己調節系が経験を通じて内部の補償反応を段階的に獲得していく過程を検討するため、本研究では、ホメオスタシスの予期的制御の数理モデルを構築した。本モデルは、予測可能な外乱が繰り返し内部の生理変数を攪乱するが、個体が試行を重ねるにつれて、外乱が生じる前に予期的に補償反応を生成するよう学習する挙動を表現することを目的とする。本モデルの中心の特徴は、予期制御を連続値の信号として表現し、その強度を試行ごとに漸増的に更新する点にある。この定式化により、生体において観察される耐性の漸進的獲得を再現できる。

前節で述べた「反応的制御」と「予期的制御」の区別に整合するよう、基礎的なホメオスタシス動態である「反応的調節」については学習の過程を通して不変とし、経験によって選択的に変化するのは、外乱に先行して動員される予期成分のみとした。以下では、まず対象とする現象の代表例の実験の知見を紹介する。そして、それを表現するための提案モデルについて、順を追って示す。はじめに内部状態と補償反応を支配する力学方程式を示し、次にドライブ関数と報酬の定義を与える。続いて、抑制学習と文脈依存的ゲーティング機構を導入し、最後に、自己誘発的な調節と外因性の攪乱とを分離することで学習を安定化させる学習則および報酬帰属機構を記述する。

2.1 行動パラダイムと対象現象

本研究では自己調節系の例として、古典的条件づけにおけるエタノール誘発性低体温課題を扱う (Mansfield & Cunningham, 1980)。Mansfield と Cunningham (1980) は、2 種類の環境を用意し、それぞれでエタノールあるいは生理食塩水を注射した。エタノール注射後には約 1.5°C の低体温が生じる (図 2A) が、試行を重ねるにつれ、動物はエタノールの効果を予期して、エタノールと結びついた環境で補償的な加温反応を示すようになり、最終的に耐性を獲得する (図 2B)。

この課題では、2 つの環境はそれ自体では快・不快を伴わない中性刺激であり、条件刺激 (conditioned stimulus; CS) として扱える。環境 1 (CS1) はエタノール注射 (unconditioned stimulus; US1) と、環境 2 (CS2) は生理食塩水注射 (US2) と対提示され、それぞれの随伴性が学習される。学習過程は、CS-US 対提示を繰り返す獲得 (acquisition)、CS は提示するが対応する US を提示しない消去 (extinction)、消去後に再び CS-US 対提示を再開する再獲得 (reacquisition) の 3 段階からなる。図 2B の消去 (E1、網掛け) では、本来 US1 と結びついた CS1 (エタノール注射環境) に対して US2 (生理食塩水) が提示されている。この試行で平時より高い体温が観察されたことは、CS1 が引き金となり、エタノール誘発性低体温を見越した補償的な加温反応が表出したことを示唆する。さらに再獲得 (R1) では、初回試行よりも低体温が緩和されていることから、獲得された耐性が消去によって完全には失われていないことが示される。

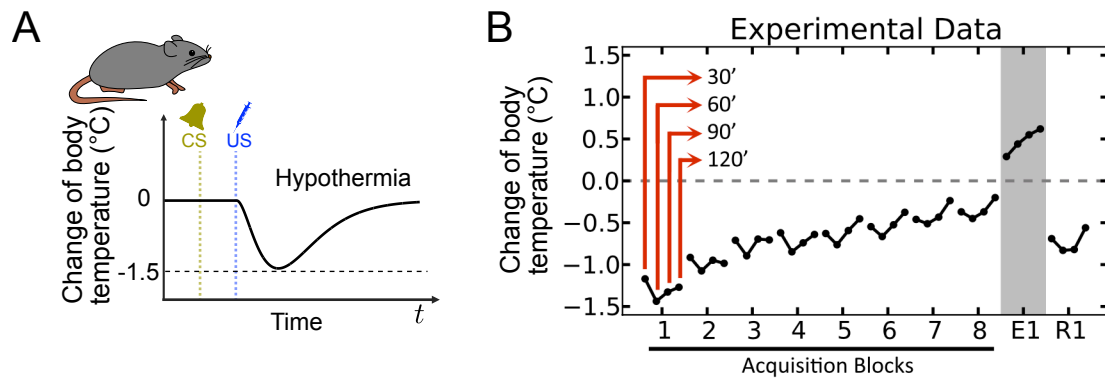


図 2.古典的条件づけにおけるエタノール誘発性低体温に対する耐性獲得 (A) 本課題における条件づけスキーム。今回は、ある環境 (CS) でエタノール注射 (US) を打つという CS と US の対提示の実験を扱う。通常、エタノール注射を実施すると 1.5 度程度の低体温が引き起こされる。(B)実験データ。実験では、エタノール (E1 では生理食塩水) 注射後 30, 60, 90, 120 分時点の体温を計測した。直腸マウスは試行を重ねるにつれて、エタノール誘発性低体温に対する耐性を徐々に獲得した。また、この耐性は消去試行における CS 提示 (第 8 試行後の E1, 網掛け部分) によって示されている。第 10 試行の R1 は、再獲得 (reacquisition) の最初の試行を表す。(Keramati and Gutkin, 2014 より転載. Mansfield & Cunningham, 1980 を改変.)

2.2 状態ダイナミクスと予期的制御

本研究では、古典的条件づけにおける体温調節に代表される自己調節系の内部状態の時間発展を、「反応的ホメオスタシス制御」と「学習によって獲得される予期的制御」を明示的に分離した連続時間力学系として記述した。 i 回目試行における時刻 t の内部状態（例：体温）を $x(t)$ とすると、そのダイナミクスは

$$\frac{dx}{dt} = -\frac{dD}{dx} + f(t - t_{US,i}) + s_i g(t - t_{CS,i})$$

で与えられる。本式は機能的に異なる 3 成分から構成される：(i) 反応的ホメオスタシス動態 dD/dx 、(ii) 外乱 $f(t)$ 、(iii) 予期的補償応答 $g(t)$ である。第 1 項 dD/dx は、内部状態を望ましい基準値 x^* へ回復させようとする反応的ドライブを表す。本モデルではドライブ関数を二乗偏差コスト

$$D(x) = a(x - x^*)^2$$

として定義し、 dD/dx が基礎的ホメオスタシス回復を保証する安定化力として働くようにした。重要な点として、この反応的機構は学習によって変更されず、全てのシミュレーションを通して常に存在する。外乱と予期的制御が存在しない場合、本項のみが基準値への指数緩和を生む。

一般性の担保のため、本研究ではホメオスタシスの基準値を $x^* = 0$ に固定し、内部状態 $x(t)$ は基準値からの偏差を表すものとした。第 2 項 $f(t - t_{US,i})$ は、US により誘発される外因性攪乱（例：エタノール誘発性低体温）を表す（図 3 左中、式は補遺 A1.1）。第 3 項 $s_i g(t - t_{CS,i})$ は、CS によって惹起される予期反応を表す。ここで $g(t)$ は CS 開始後に生じる補償応答の典型的な時間プロファイルを表し、その振幅が試行依存のスカラー s_i によって調節される（図 3 左下、全モデル式は補遺 A1.1）。ここで重要なのは、モデルにおける学習が s_i の更新のみによって進行する点である。すなわち、エージェントは経験により、この予期成分の強度を徐々に調整し、外乱による影響が顕在化する前にそれを相殺するように学習する。この構造により、基礎的なホメオスタシス安定性を維持したまま、経験依存的に補償反応が段階的に獲得される過程を表現できる。

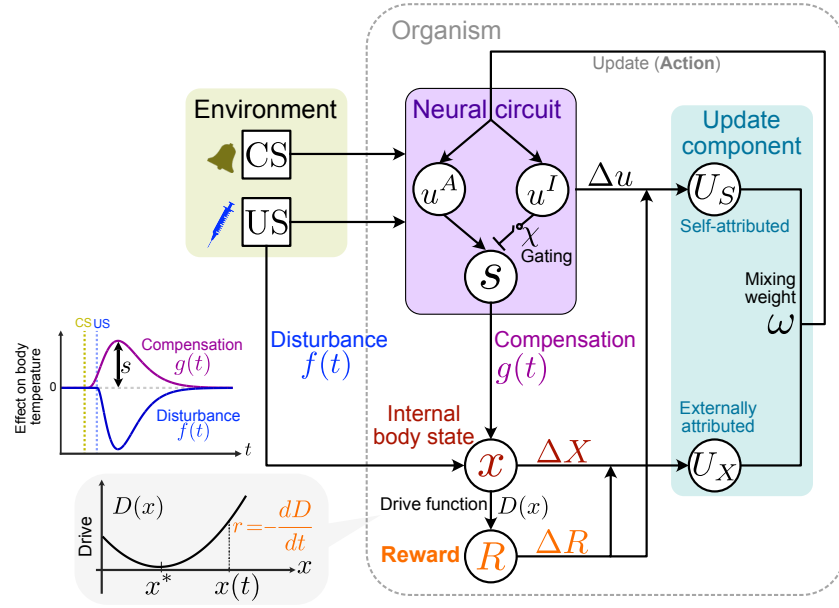


図 3. 本研究で提案するホメオスタシス強化学習モデルの模式図。環境と個体の相互作用を示す。左中段：内部状態生成の概念図。CS の開始に伴い、補償信号 $g(t)$ が誘発され、US 開始で生じる外乱 $f(t)$ を相殺するように作用する。補償信号 $g(t)$ の振幅 s は、活性化および抑制の重み u^A 、 u^I により定まる。左下段：瞬時報酬 $r(t)$ は、現在の状態におけるドライブ関数 $D(x)$ の時間微分として定義されるが、学習に直接働くのは、瞬時報酬をまとめて1試行あたりの報酬量として計算した R である。具体的には、試行間の報酬変化 ΔR について、自己生成的な制御更新 Δu に起因する成分 (U_S) と、外因性に誘発された状態偏差 ΔX に起因する成分 (U_X) に帰属される。これらを混合重み ω により統合し、制御パラメータ u^A および u^I を更新する。

2.3 ドライブ低減と報酬の定義

上で導入したドライブ関数 $D(x)$ に基づき、本モデルの学習は、内部状態がホメオスタシスの目標値に対してどのように変化したかを反映する瞬時報酬信号によって駆動される。具体的に、報酬はドライブ低減の時間率として

$$r(t) = -\frac{dD}{dt}$$

と定義する。この定義の下では、基準値 x^* からの偏差を減少させる変化は正の報酬を生み、偏差を増大させる変化は罰（負の報酬）となる。さらに、生体調節においてしばしば観察される非対称性、すなわち悪化側への偏差が同程度の改善よりも重大になりやすい性質を表現するため、次のように負の報酬に対して非対称な重み付けを導入する。

$$r(t) = \begin{cases} -\frac{dD}{dt} & \text{if } \frac{dD}{dt} \leq 0, \\ -\eta \frac{dD}{dt} & \text{if } \frac{dD}{dt} > 0. \end{cases}$$

瞬時報酬は各試行内で積分され、試行 i における総報酬 R_i を与える。これが予期的制御の更新に用いられる学習信号となる。

2.4 予期信号・抑制学習・ゲーティング機構

試行 i において CS が提示された際に動員される補償応答の振幅は、予期制御信号 s_i によって決まる。形式的には

$$s_i = A_i - \chi_i w^{A,I} I_i$$

と定義する。ここで A_i および I_i はそれぞれ活性化（activation）ユニットと抑制（inhibition）ユニットを表し、 $w^{A,I} > 0$ は抑制影響の強さ、 χ_i は文脈依存的ゲーティング変数である。すなわち s_i は、予期補償を促進する活性化経路と、それを抑圧する抑制経路という、拮抗する 2 成分の相互作用として生成される。活性化・抑制の両ユニットは CS 入力によって駆動され、

$$A_i = u_i^A \text{CS}_i, \quad I_i = u_i^I \text{CS}_i$$

と表される。ここで u_i^A と u_i^I は試行依存の重みであり、学習によって更新される。したがって学習は、予期的信号に対する興奮性寄与と抑制性寄与のバランスを調整することとして実装される。

ゲーティング変数 $\chi_i \in \{0,1\}$ は、抑制影響が発現するか否かを制御する。直前試行における CS-US 随伴性に基づき

$$\chi_i = 1 - \text{CS}_{i-1} \text{US}_{i-1} \quad (\text{CS}_{i-1}, \text{US}_{i-1} \in \{0,1\})$$

と定義する。直前試行で CS-US の対提示が維持されていた場合には $\chi_i = 0$ となり抑制は抑え込まれる。一方、随伴性が破綻した場合（例：消去；extinction）には $\chi_i = 1$ となり抑制影響が解放される。なお χ_i は直前試行に依存するため、消去の最初の 1 試行目（図 4, 試行 9）では抑制発現はまだ誘発されず、随伴性違反を経験した後（図 4, 試行 10）に顕在化する。この定式化により、元の興奮性連合を消去するのではなく、抑制学習を追加する形で消去学習を実装できる。

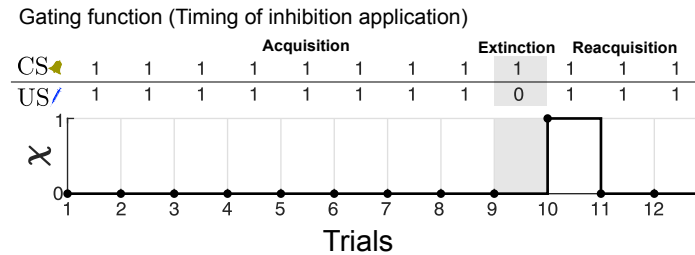


図 4. CS-US に基づくゲーティング変数 χ の挙動. 各試行における CS および US の提示有無（上段）に応じて、ゲーティング変数 χ が更新される様子を模式的に示す。獲得期および再獲得期（消去期直後の 1 試行を除く）では CS と US が同時に提示されるため $\chi = 0$ となり抑制成分（ I から A への抑制）の寄与が遮断される。一方、消去期（消去期の 1 試行目を除く）では CS のみが提示されるため $\chi = 1$ となり抑制が有効化される。

2.5 学習則と報酬帰属

本モデルの学習は試行単位で進行し、総報酬を最大化するように予期制御信号 s_i を調整することを目的とする。学習するエージェントは連続する試行間の報酬変化を評価し、それが自己生成的な予期制御によるものか、あるいは外因性の攪乱による内部状態の変化によるものを区別して帰属する。

試行 i の総報酬は瞬時報酬の時間積分として

$$R_i = \int r(t) dt$$

と定義される。学習は試行間の報酬差

$$\Delta R_i = R_i - R_{i-1}$$

によって駆動され、これは試行レベルの報酬予測誤差として機能する（完全な更新式は補遺 A1.1）。ここでの重要な点として、 ΔR_i の変化は質的に異なる二種類の要因から生じ得る。第一に、予期制御信号 s_i の調整が成功し、報酬が漸進的に改善する場合である。第二に、CS-US 随伴性の変更などにより、エージェント自身の制御では説明できない形で内部状態が予期せず変化し、その結果として報酬が急変する場合である。そこで我々は、 ΔR_i の帰属曖昧性を解消するため、学習信号を自己帰属成分 U_S と外因帰属成分 U_X に分解した。自己帰属成分 U_S は、報酬変化が制御パラメータ $u_i^{(m)}$ の変化とどの程度共変するかを定量化し、

$$U_S^{(m)} = h \left(\left| \frac{\Delta R_i}{\Delta u_i^{(m)} + \varepsilon} \right| \right), \quad m \in \{A, I\}$$

で与える。ここで m は活性化(A)または抑制(I)を表し、 $h(\cdot)$ は更新量を有界化して数値安定性を保証する飽和型非線形関数である（補遺 A1.1, 図 S1A）。この項は自己調節による漸進的な報酬改善を捉える。一方、外因帰属成分は内部状態の偏差 X_i に関連する報酬変化を捉え、

$$U_X^{(m)} = h \left(\left| \frac{\Delta R_i}{\Delta X_i + \varepsilon} \right| \right)$$

と定義する。これは、報酬変化が制御信号の微小更新からは説明できない場合に支配的となる。両者の相対寄与は混合重み $\omega_i \in [0,1]$ により決定し、

$$\omega_i = \sigma \left(h \left(\left| \frac{\Delta R_i}{\Delta s_i + \varepsilon} \right| \right) - c \right)$$

とする。ここで $\sigma(\cdot)$ はシグモイド関数、 c は閾値パラメータである（閾値の設定については図 S1B, 閾値の感度解析については補遺 A1.3 参照）。報酬変化が s_i の漸進的調整と整合的であるとき ω_i は小さく、学習は主として U_S によって駆動される。逆に、自己誘発的な制御変化に比して報酬変化が不釣り合いに大きい場合には ω_i が増大し、学習は U_X 側に移行する。これにより、環境由来の急激な攪乱を自己の行為に誤って帰属することが抑制され、学習が

安定化する。制御パラメータ u_i^A および u_i^I は、これら 2 成分の加重結合として更新される（補遺 A1.1.1）。

総じて、本帰属機構は、試行レベルのモデルフリーな報酬最大化を可能にしつつ、突発的攪乱や文脈変化に対して頑健な学習を実現する。

2.6 シミュレーション手順と評価指標

次に、本章で定義した一次元モデル（全式は 補遺 A1.1）を用いて、古典的条件づけ試行列のシミュレーションを行った。各試行では、条件刺激（CS）の提示開始を起点として内部状態 $x(t)$ の連続時間ダイナミクスを發展させ、各試行の終了時刻は、刺激提示開始（CS 開始）から一定の観測窓 T_{trial} を設定し、 $t = T_{\text{trial}}$ で一律に打ち切った。

シミュレーションは全 12 試行から構成した。具体的には、予期的補償反応の獲得を促すための CS-US 対提示による条件づけ獲得試行（試行 1–8）、学習された随伴性を破綻させるために CS のみを提示する消去試行（試行 9）、および CS-US 対提示を再開して学習反応の回復を評価する再獲得試行（試行 10–12）である（図 4 上段参照）。

各モデル要素の機能的寄与を明確にするため、4 つの条件で比較検討を行った（図 5B–F）。

1. **反応的ホメオスタシス動態のみ**（予期的調節なし）：学習された予期制御の寄与を切り分けるため、予期項 $s_i g(t)$ を除去した条件をシミュレートした。この場合、系は反応的ホメオスタシス動態と外乱項のみに基づいて挙動する（図 5A）。
2. **ゲーティング機構の有無**：消去および再獲得における文脈依存的ゲーティングの役割を検討するため、ゲーティング変数 χ_i を有効化した条件と無効化した条件を比較した。後者では抑制学習自体は存在するものの、随伴性違反に応じて抑制の発現を選択的に解放できない（図 5B–C）。
3. **報酬の対称性／非対称性**：損失優位（loss-dominant）な価値付けの影響を検討するため、対称報酬条件（ $\eta = 1$ ）と、基準値からの悪化（逸脱の増大）を改善より強く重み付けする非対称報酬条件（ $\eta > 1$ ）を比較した（図 5B, D）。
4. **基準 HRL モデルとの比較**：参照として、既報のホメオスタシス強化学習モデル（Keramati & Gutkin, 2014）をシミュレートした。このモデルは価値に基づく強化学習により離散的な補償行動を選択する点で、本研究の連続的な予期信号の表現・更新とは異なる。一方で、状態ダイナミクスと報酬定義は本研究モデルと同一とした（図 5E；実装の詳細は 補遺 A1.2）。

学習の程度は、主として以下の観点から評価した：(i) 試行を通じた $x(t)$ の変化の漸進性、(ii) 消去後の総報酬 R_i の回復、(iii) 再獲得期における予期反応の再発現。特に断りのない限り、各条件間でパラメータは共通とし、条件差は当該要素の有無または設定のみに由来するよう統制した（補遺, 表 S2）。パラメータの感度解析は補遺 A1.3 を参照されたい。

本研究は MATLAB 2024a を用いて行い、コードは github にて公開予定である。数値積分について、連続時間の力学系は、MATLAB の適応ステップ ODE ソルバ ode45（陽的 Runge–Kutta 法（4,5））を用いた。相対許容誤差および絶対許容誤差はそれぞれ 10^{-6} 、 10^{-9} に設定した。

第三章

結果Ⅰ（一次元モデル）

本章では、第二章で記述した一次元モデルを用い、条件づけ・消去・再獲得からなる試行列をシミュレートし、予期的調節がどのように段階的に獲得され、抑制され、そして再発現するかを検討した。次に、主要なモデル要素の機能的寄与を明らかにするため、(i) 予期的調節の除去、(ii) ゲーティング機構の有無、(iii) 報酬重み付けの対称性/非対称性、(iv) 既存 HRL モデルとの比較、の 4 つの操作条件を設定した。モデル性能の評価は、上記にて内部状態 $x(t)$ と試行総報酬 R_i の試行ごとの軌跡に基づいて行ったが（図 5）、主要なモデル要素の機能的寄与についてさらなる考察のため、モデルシステムを構成する変数についても検討した（図 6-9）。

3.1 反応的ホメオスタシス動態のみでの挙動

まず、予期的調節を無効化した条件として、予期的制御項 $s_i g(t)$ を除去し、反応的ホメオスタシス動態のみを残して系の挙動を調べた。この条件では、各外乱の後に内部状態は一貫して基準値へ回復し、基礎的なホメオスタシス回復機構が保たれていることが確認された（図 5A）。一方で、試行を重ねても耐性は獲得されず、CS-US の反復対提示によっても、外乱によって生じる偏差の大きさは変化しなかった。したがって、反応的ホメオスタシス動態は攪乱後の安定性回復には十分であるが、予期的補償や漸進的な耐性獲得を成立させるには不十分である。本条件は、予期的制御の寄与を切り分けるための対照条件であり、耐性獲得が生じないという結果はモデル設計の意図と一致した。

3.2 予期的耐性の漸進的獲得

次に予期的調節を有効化すると、条件づけ試行を通じて耐性が段階的に獲得された（図 5B, C, 試行 1-8）。CS-US 対提示を反復するにつれ、CS 開始時に補償反応を予期的に立ち上げる学習が進行し（図 6A, F, 7A, F）、低体温による偏差の大きさは試行ごとに漸減した。この試行間の偏差低減は、経験依存的に予期制御が強化されることを示しており、条件づけにより体温調節性耐性が形成される実験所見（図 2B）と整合的である。重要な点として、この漸進的耐性獲得は、ゲーティング機構の有無に依存せず（図 5B, C）、ゲーティングは予期的調節の「獲得」自体には必須ではないことが示された。これは、ゲーティング機構が、消去試行（試行 9）を経験後に初めて機能する機構であるため、それまでは機構の有無で異なる挙動をしないことによる。図 6（ゲーティングあり）と図 7（ゲーティングなし）の内部パラメータの挙動を比較しても、ゲーティング機構が機能した結果を反映する試行 11 までは、全く同じ挙動が観察された。

一方で、報酬と罰の重み付けを対称（ $\eta = 1$ ）とした場合には、ゲーティング機構が存在しても学習が成立しなかった（図 5D 左）。この条件では、 $\eta = 1$ としたことにより、各試行内の改善と悪化が相殺され、報酬がほぼ 0 となる（図 5D 右）。そのため、報酬により駆動される強化学習である本モデルでは、学習信号が消失することとなる。内部パラメータの挙動も体温調節性耐性として十分な補償反応を立ち上げるまで学習が進まないことを示した（図 8G, 8F）。以上より、効果的な予期的調節を漸進的に獲得させるには、悪化側を強く評価する非対称的価値付けが本質的であることが示された。

3.3 ゲーティングによる消去と迅速な再獲得

続いて、消去および再獲得期の挙動を比較し、ゲーティング機構の寄与を検討した（図 5B と図 5C）。消去試行（CS のみ提示）後、ゲーティングあり・なしの両機構で、補償反応は同程度に抑制されたが、その後 CS-US 対提示を再開すると、ゲーティングを備えたモデルでは、予期反応が直ちに回復し、それに伴って報酬が急増した（図 5B、試行 11）。これに対し、文脈依存的ゲーティング χ を除去した条件では、活性化・抑制ユニットが存在していても同様の迅速な回復は見られなかった（図 5C、試行 11, 図 7A, 7F）。以上は、実験で観測された応答の迅速な再獲得、ここでの抑制された予期反応の再発現には、ゲーティングが不可欠であることを支持する結果となった。このときの内部変数の挙動については図 6（ゲーティングあり）と図 7（ゲーティングなし）を参照されたい。

3.4 既存 HRL モデルとの比較

最後に、予期反応を離散的な行動集合から選択する既存の HRL 定式化（Keramati and Gutkin, 2014）と比較した。この基準モデルでは、耐性の漸進的獲得が再現できず（図 5E 左）、試行ごとの報酬も不安定であった（図 5E 右）。図 9 に示したように、価値観数 Q （取り得る行動の価値を意味し、これに基づいて次の行動を選択する）の意味する耐性を発揮する、しないという行動の価値の差は学習を経て徐々に開き（図 9D）、この二択の選択確率も耐性を発揮する方が徐々に高くなっていった（図 9E）。しかしながら、例えば選択確率の変化が漸進的であったとしても、予期反応は離散的な行動集合からの選択になるため、体温調節耐性の獲得自体は漸進的にはならない。この結果は、段階的で経験依存的な調節を表現するうえで、連続値としての予期制御信号が重要であることを示唆した。

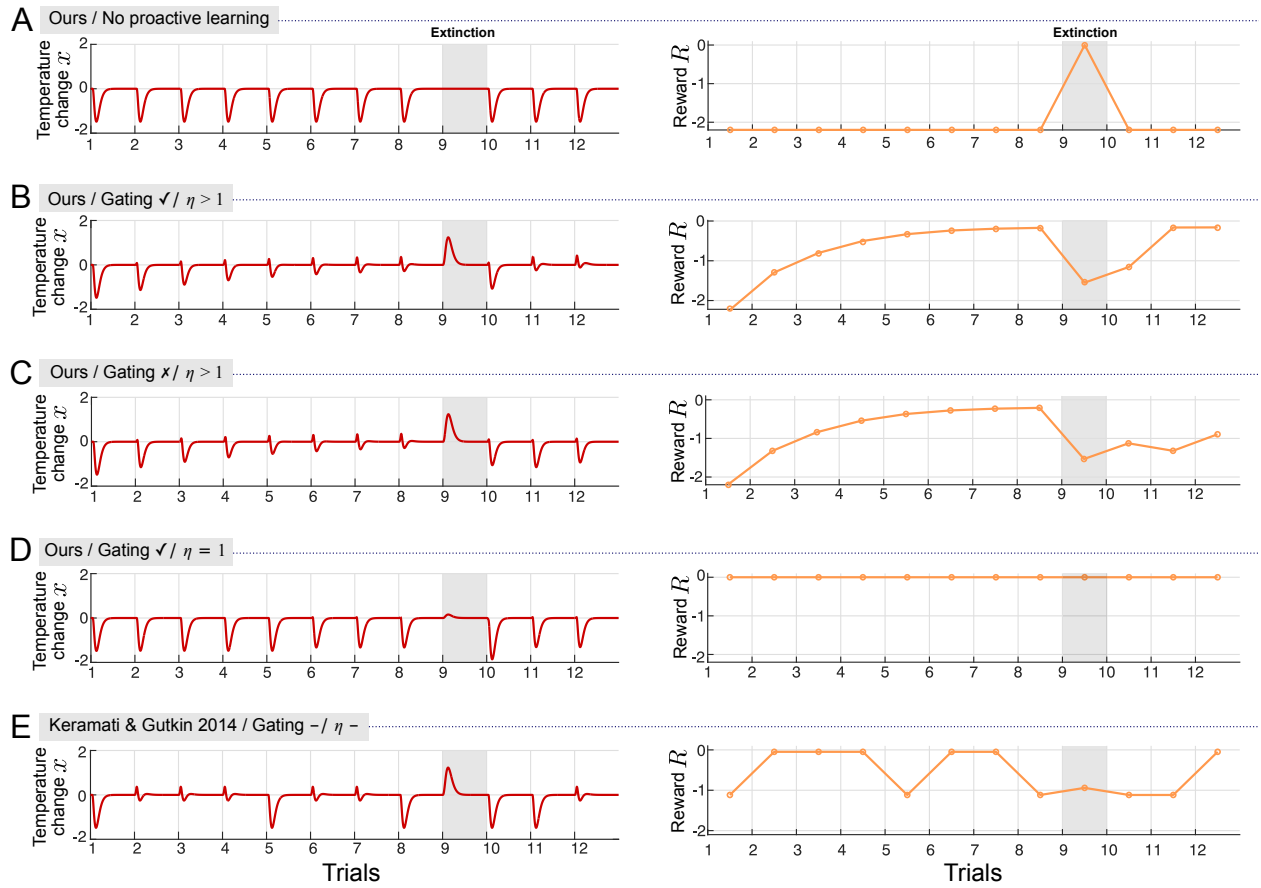


図 5. 一次元モデルによる体温のホメオスタシス調節 (A-E) 一次元モデルでシミュレートした体温変化 (左列) と、各試行で得られる報酬 (右列)。(A) 予期的補償成分を除いた提案モデル。(B) ゲーティング関数 χ を導入し、再獲得期に抑制を抑圧する条件。(C) ゲーティング機構を除いた条件。(D) ゲーティングは導入するが、報酬構造を対称($\eta = 1$)とした条件。(E) 離散的な「全か無か」型の補償反応を選択する既存 HRL モデルによる結果。A、B、C、および E では報酬の重み付けパラメータを $\eta = 1.5$ に設定した。網掛け領域は消去試行を示す。

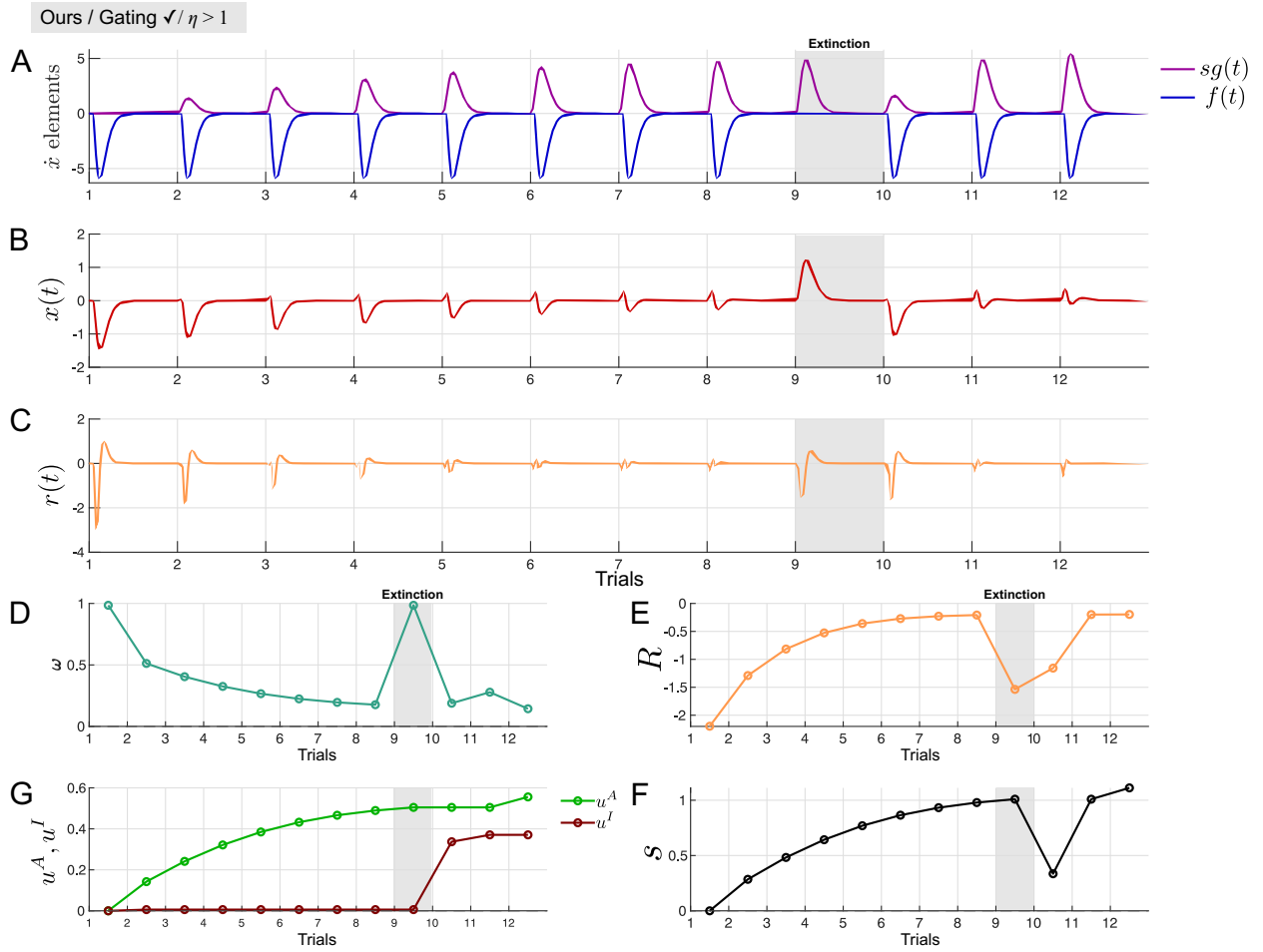


図 6. 一次元モデルのパラメータ挙動：ゲーティングあり・非対称報酬重み付け条件 ($\eta > 1$). (A) 外乱 $f(t)$ と、学習された予期的補償反応 $sg(t)$ の時系列。網掛け部分は消去学習の試行を示す。(B) 内部状態 $x(t)$ のダイナミクス。(C) ドライブ関数の時間変化から導出した瞬時報酬 $r(t)$ 。(D) 偏差量の累積指標の試行ごとの推移。(E) 各試行で得られた報酬 R_i 。(F) 予期信号 s_i の試行間推移。(G) 興奮性および抑制性の重み (u_i^A , u_i^I) の試行ごとの推移。網掛け領域は消去試行を示す。本図は、図 5. B に示した結果を出力するダイナミクスの詳細を示す参照図である。

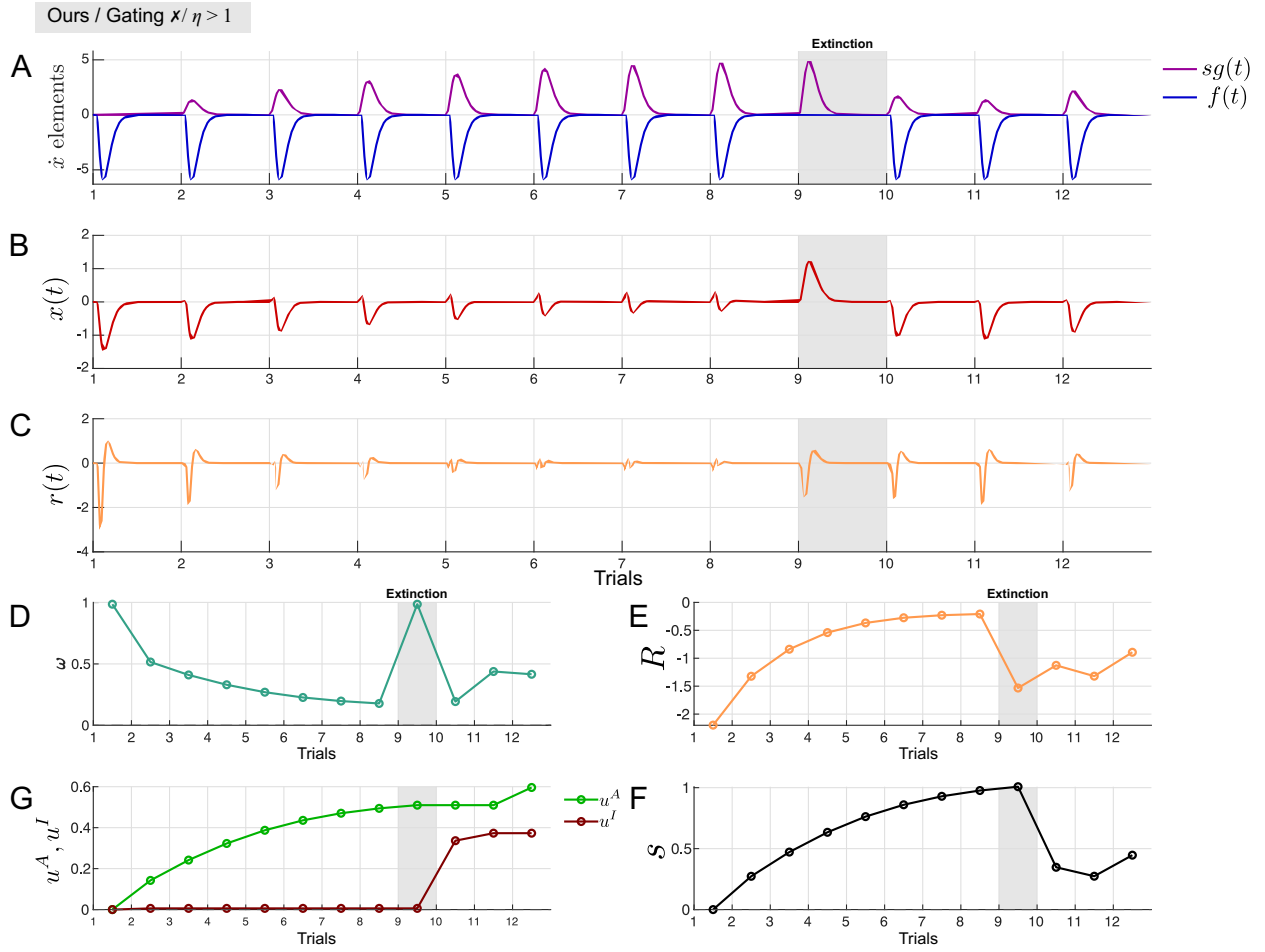


図 7. 一次元モデルのパラメータ挙動：提案モデル・ゲーティングなし・非対称報酬重み付け条件 ($\eta > 1$). (A) 外乱 $f(t)$ と、学習された予期的補償反応 $sg(t)$ の時系列。網掛け部分は消去学習の試行を示す。(B) 内部状態 $x(t)$ のダイナミクス。(C) ドライブ関数の時間変化から導出した瞬時報酬 $r(t)$ 。(D) 偏差量の累積指標の試行ごとの推移。(E) 各試行で得られた報酬 R_i 。(F) 期待信号 s_i の試行間推移。(G) 興奮性および抑制性の重み (u_i^A , u_i^I) の試行ごとの推移。網掛け領域は消去試行を示す。本図は、図 5. C に示した結果を出力するダイナミクスの詳細を示す参照図である。

Ours / Gating $\checkmark / \eta = 1$

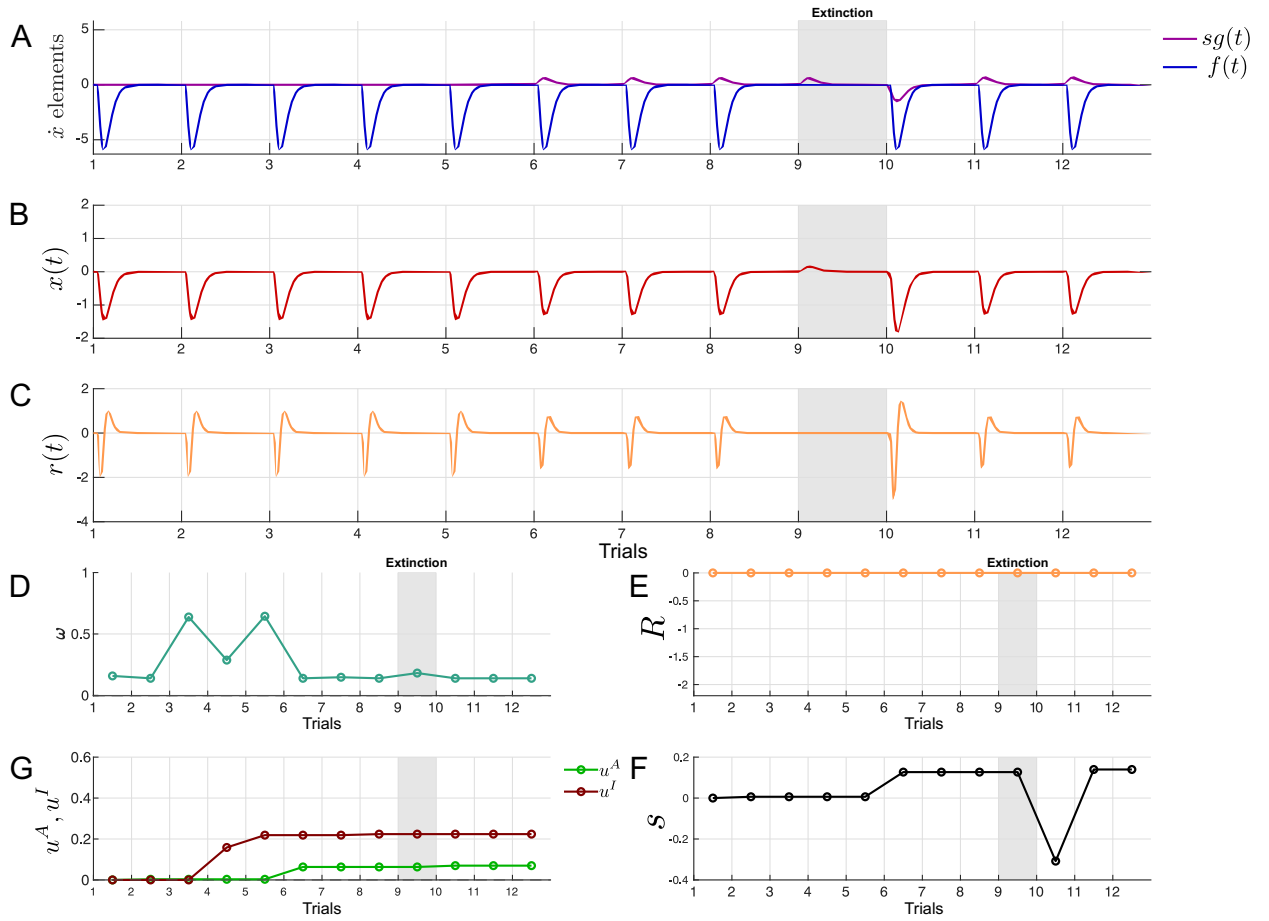


図 8. 一次元モデルのパラメータ挙動：提案モデル・ゲーティングあり・対称報酬重み付け条件 ($\eta = 1$). (A) 外乱 $f(t)$ と、学習された予期的補償反応 $sg(t)$ の時系列。網掛け部分は消去学習の試行を示す。(B) 内部状態 $x(t)$ のダイナミクス。(C) ドライブ関数の時間変化から導出した瞬時報酬 $r(t)$ 。(D) 偏差量の累積指標の試行ごとの推移。(E) 各試行で得られた報酬 R_i 。(F) 予期信号 s_i の試行間推移。(G) 興奮性および抑制性の重み (u_i^A , u_i^I) の試行ごとの推移。網掛け領域は消去試行を示す。本図は、図 5. D に示した結果を出力するダイナミクスの詳細を示す参照図である。

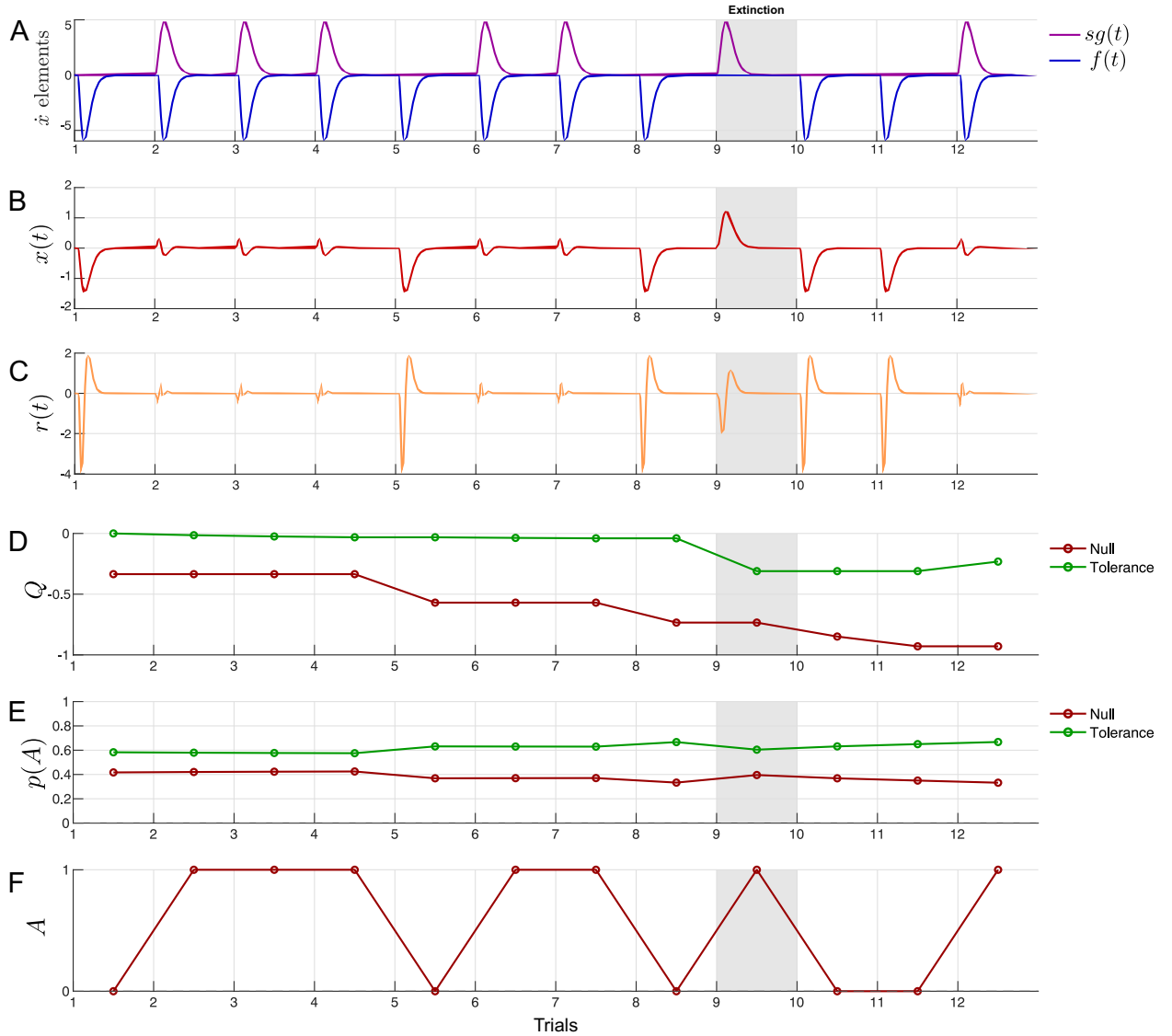


図 9. 一次元モデルのパラメータ挙動：既存モデル・Keramati and Gutkin (2014) に基づく基準 HRL モデル (A) 外乱 $f(t)$ と、学習された予期的補償反応 $sg(t)$ の時系列。網掛け部分は消去学習の試行を示す。(B) 内部状態 $x(t)$ のダイナミクス。(C) ドライブ関数の時間変化から導出した瞬時報酬 $r(t)$ 。(D) 偏差量の累積指標の試行ごとの推移。(E) 各試行で得られた報酬 R_i 。(F) 予期信号 s_i の試行間推移。(G) 興奮性および抑制性の重み (u_i^A , u_i^I) の試行ごとの推移。網掛け領域は消去試行を示す。本図は、図 5.E に示した結果を出力するダイナミクスの詳細を示す参照図である。

第四章

モデルⅡ

共有メタ信号による予期的制御
(多次元拡張モデル)

前章までの一次元モデルは、体温という単一の内部変数をホメオスタシス強化学習の枠組みで調節することに焦点を当てた。しかし実際の生体は、多数の内部変数を同時に一定範囲へ保つ必要があり、しかもそれらはしばしば共通の調節資源（制約）を共有している。例えば、交感神経緊張、バソプレシン、コルチゾールなどの調節信号が、体温・血圧・水分状態・血糖など複数の生理変数へ同時に影響を及ぼし得る。これにより、ある変数を最適域へ近づけようとする調整が別の変数を意図せず攪乱することは避けられず、調節には本質的なトレードオフが生じる。このような生理制御の本質的特徴を捉えるため、本研究ではモデルを多次元へ拡張し、複数の内部変数が、より少数の共有制御信号（本研究ではメタ信号と呼ぶ）によって同時に調節される枠組みを導入した（図 10A）。なお、ホルモンや自律神経活動は「状態」と「制御入力」の両方として扱い得るため、生理学的に絶対的な線引きは一意ではない。本研究では制御理論の慣例に従い、目標範囲への維持が要請される量を「生理変数」、それらを調節するために動員される制約を伴う量を「調節資源」と定義する。ホルモンについては、必要に応じて濃度を状態、分泌を入力として分離して表現できるが、本章では抽象化のため信号として扱う。

本モデルでは、「各メタ信号が固定の影響重み W を介して複数の内部変数に作用する」ことで調節資源の制約を表現した。したがって有効な調節とは、各変数を独立に最適化することではなく、相反するホメオスタシス要求の間で折り合いをつける過程として定式化される。直感を得るため、まず最も単純でありながら自明ではない設定として、1 つのメタ信号（ $K=1$ ）が 2 つの内部変数（ $N=2$ ）を制御する場合を扱う。この縮約系は、単一の共有制御が異なる内部変数に対して（多くの場合は反対方向を含む）影響を同時に及ぼし得るという共有制御の本質を保持しつつ、解析的・視覚的に扱いやすい。まずこの設定で多次元シミュレーション（図 10）を行い、その後、より一般的な複数メタ信号・複数変数の場合（図 11）へ拡張する。

4.1 共有メタ信号・影響重み・優先度

拡張モデルは、 N 個の内部変数

$$\mathbf{x} = (x_1, x_2, \dots, x_N)$$

からなり、各変数 x_n は固有の基準値（セットポイント） x_n^* をもつ。一次元モデルと異なり、調節は各変数に独立な制御信号によって行うのではなく、 $K(< N)$ 個の共有メタ信号

$$\mathbf{z} = (z_1, z_2, \dots, z_K)$$

によって媒介される。各内部変数 x_n は局所制御信号 s_n により調節され、 s_n はメタ信号の加重和として

$$s_n = \sum_{k=1}^K w_{kn} z_k$$

と定義する。ここで w_{kn} は、メタ信号 z_k が変数 x_n に及ぼす生理学的影響を表す固定パラメータである（図 10A）。影響重み w_{kn} は正負のいずれも取り得るものとし、共有調節信号が標的ごとに異質、場合によっては相反する効果をもつ状況を表現する。結果として、ある

メタ信号を増加させて一方の変数を安定化させることが、同時に別の変数を不安定化させる可能性が生じる。これは例えば、発汗による蒸発冷却が体温を低下させる一方で体内水分を失わせる、といった現象や、薬理介入が標的以外の生理次元を攪乱して副作用を生む、といった状況に対応する。

各内部変数には二次形式のドライブ関数

$$D_n(x_n) = a_n(x_n - x_n^*)^2$$

を割り当て、係数 a_n を変数 x_n に与えられた優先度として解釈する。多次元モデルにおける学習はメタ信号レベルで行われ、メタ信号 z_k を規定する制御変数が、全変数で集約した総報酬

$$R_{\text{tot}} = \sum_n R_n$$

を最大化するよう更新される。各メタ信号は複数の変数へ異なる優先度で作用するため、メタ信号の最適化は必然的に競合する要求のバランス調整となり、優先度の高い変数の安定化を優先しつつ、優先度の低い変数の逸脱を一定程度許容する解として現れると予想される。多次元ダイナミクス・報酬・学習則の厳密な定義は補遺 A2 にまとめた。

4.2 シミュレーション手順と評価指標

4.1 で述べた多次元モデル（全式は補遺 A2）を用い、共有メタ信号が複数内部変数をどのように調節し、学習がそれらのトレードオフをいかに折衝するかを検討した。基本的な試行構成は一次元モデルと同様としつつ、多次元制御に特有の解析を追加した。

4.2.1 単純設定：1 つのメタ信号が 2 変数を制御する場合

共有メタ信号による制御の直感的理解を得るため、まず $K = 1$ 、 $N = 2$ の単純非自明系を扱った（図 10 A）。この設定では、単一の制御信号が異なる変数に対して相反する効果を同時に及ぼし得る、という共有調節の本質が最も明瞭に現れる。メタ信号の影響は固定行列

$$W = (1, -1)$$

で与え、メタ信号が 2 つの局所制御信号へ正負逆符号で写像されるようにした。この選択は特定の生理経路を指すものではなく、「一方を改善すれば他方が必ず悪化する」という双方向の影響の一般例を与えることを意図した。

具体例として、外乱をエタノール誘発性低体温に対応づけ、2 つの変数を体温と抽象的な燃料関連状態（例：Adenosine triphosphate, ATP, 代謝予備能）として解釈した。すなわち、低体温を防ぐ熱産生のためにエネルギーを消費する例を考えた。試行列は一次元モデルと同一の 12 試行（条件づけ 1-8、消去 9、再獲得 10-12）を適用し、内部変数・メタ信号・報酬のダイナミクスを図 10B-G に示した。優先度は $\mathbf{a} = (2, 2)$ に固定した。

4.2.2 拡張設定：2 つのメタ信号が 3 変数を制御する場合

続いて、より複雑なトレードオフを検討するため、 $K = 2$ 、 $N = 3$ の系をシミュレートした（図 12-16）。影響行列は

$$W = \begin{pmatrix} 1 & -1 & 0 \\ 0 & 1 & -1 \end{pmatrix}$$

とした。全てのシミュレーションにおいて、外乱は第 1 変数 x_1 のみに直接作用し、残りの変数は共有メタ信号を介して間接的に影響を受ける（図 12A）。以下の 2 点に焦点を当てて多次元調節を解析した。

(A) 局所変数間の優先度付けの効果

各内部変数 x_n に優先度重み a_n を割り当て、全体ホメオスタシスに対する相対的重要度を表現した。外乱は常に x_1 に同一に与えたまま、優先度パターン $\mathbf{a} = (a_1, a_2, a_3)$ を変化させたときの時系列挙動を比較した。評価には、基準値からの偏差 $x_n(t) - x_n^*$ の軌跡と総報酬 R_{tot} を用い、優先度設定がトレードオフと回復ダイナミクスをどのように規定するかを検討した。

(B) 報酬非対称性と優先度の同時効果

学習性能をパラメータ空間全体で系統的に特徴づけるため、報酬の非対称性 η と優先度パターン θ の同時効果を調べた。 η は区間 $[1, 3]$ から等間隔に 15 点をサンプルした。優先度重み $\mathbf{a}$ は、 $\theta \in [0, 2\pi)$ の 1 つの角度変数でパラメータ化し、位相をずらした余弦関数により 3 変数の相対優先度が滑らかに循環するように定義した（図 17A, 補遺 A2. これにより、比較的均衡した設定から強い不均衡設定までを、対称性を保ったまま偏りなく掃引できる。完全に均衡した設定はこれとは別にシミュレーションを行った）。そして、各パラメータ組 (η, θ) について、メタ信号空間上で総報酬 R_{tot} を二次元グリッドとして評価した。具体的には、 $z_1 \in [-2.5, 4]$, $z_2 \in [-5, 1.5]$ の範囲から 65 通りの組合せを等間隔にサンプルし、各 (η, θ) について 15×15 の格子として可視化した（図 17B）。報酬最適なメタ信号は、

$$\mathbf{z}^*(\eta, \theta) = \underset{\mathbf{z}}{\operatorname{argmax}} R_{\text{tot}}(\mathbf{z}; \eta, \theta).$$

とし、固定パラメータ下で R_{tot} を最大化する格子点として定義した。この最適点はダイナミクスの定常解ではなく、あくまで数値的参照として導入したものである。

強化学習がこの最適参照へ収束するかを評価するため、消去を含まない CS-US 試行を 100 回連続でシミュレートし、最終 10 試行における学習軌跡 $\mathbf{z}_i$ と $\mathbf{z}^*(\eta, \theta)$ のユークリッド距離の平均を収束指標として用いた。

第五章

結果Ⅱ（多次元拡張モデル）

一次元モデルでは、単一の内部変数の偏差をどれだけ効果的に低減できるかによって、予期的調節の成否を評価できた。しかし生体の調節は通常、単一の次元に限られない。むしろ複数の内部変数が同時に調節され、しかもそれらが共有された制御信号により結び付けられていることが多いため、調節には本質的に回避不能なトレードオフが伴う。

このようなトレードオフの性質について検討するため、第四章で導入した多次元モデルを用い、限られた数のメタ信号によって複数の内部変数が同時に制御される状況下で、予期的ホメオスタシス調節がどのように機能するかを検証した。この枠組みでは、ある変数の状態を改善しようとする、別の変数に偏差を誘発し得るため、学習は各変数を独立に最適化するのではなく、競合するホメオスタシス要求の間で折り合いをつけながら進む必要がある。まず、共有制御に伴うトレードオフがどのように生じるかを示すために、1つのメタ信号が2つの局所変数を制御する単純設定を解析した（図 11）。次に、2つのメタ信号と3つの局所変数からなるより高次元の系へ拡張し、「優先度」と「報酬非対称性」が、多次元にわたる予期的調節の協調において強化学習が成功するか失敗するかをどのように規定するかを系統的に検討した（図 12-16）。

5.1 共有メタ信号制御下におけるトレードオフの創発

まず、1つのメタ信号が優先度の等しい2つの局所内部変数を、互いに反対方向の影響として制御する単純な多次元設定（図 10A； $K = 1, N = 2$ ）を検討した。この設定では、メタ信号が一方の変数を正方向に調節する一方で、同時に他方の変数へ負方向の影響を及ぼす。これは、単一の生理学的制御信号が、異なるホメオスタシス次元に対して競合的效果を生む状況を模している。低体温の例に即して言えば、冷却を相殺するためにメタ信号を増加させることは、抽象的なエネルギー関連変数に必然的なコストを課すことを意味し、トレードオフは制御の失敗ではなく、共有制御に内在する帰結として現れる。

CS-US 対提示を反復すると、系は試行を通じてメタ信号を徐々に増加させるよう学習した（図 11E）。その結果、外乱が直接作用する変数 x_1 では、試行を重ねるにつれて偏差が漸減し、予期的調節が獲得されたことが示された（図 10B、橙）。しかしこの改善は、同じメタ信号が反対符号の影響を介して間接的に作用する第2変数 x_2 における系統的な偏差を伴っていた（図 11B、青）。すなわち、ある変数の予期的調節の成功は、別の変数におけるコストを不可避に引き起こし、共有制御が課す明示的なトレードオフが観察された。この構造は報酬レベルにも反映された。2変数それぞれの試行総報酬 R_n は互いに逆方向の傾向を示し、一方の改善が他方の悪化を代償として生じることが確認された（図 10C, F）。それにもかかわらず、変数間で報酬を加算した総報酬 $R_{\text{tot}} = \sum_n R_n$ は試行を通じて増加し（図 10G）、局所的な競合が存在する状況下でも、学習が系全体としての性能を最適化できることが示された。

重要な点として、メタ信号が各変数へ及ぼす影響は符号も含めて固定であり単調であるにもかかわらず、系全体に対する「有益性」は単調ではなかった。メタ信号が増大すると、 x_1 に対する追加的な補償は、 x_2 に誘発される偏差が十分に大きくなってコストが増すにつれ、次第に利得が小さくなる（飽和に類似した挙動）。この非単調性は、影響行列 $\mathbf{W}$ が状態依存的に変化した結果ではなく、複数のドライブが同時に競合することにより、固定された影響構造の下でも $z \rightarrow R_{\text{tot}}$ のレベルで自然に創発したものである。

以上より、複数の内部変数が共有メタ信号を介して制御される場合、予期学習は各偏差を独立に最小化する過程ではなく、トレードオフの下で競合する要求を調停する過程として必然的に成立することが示された。

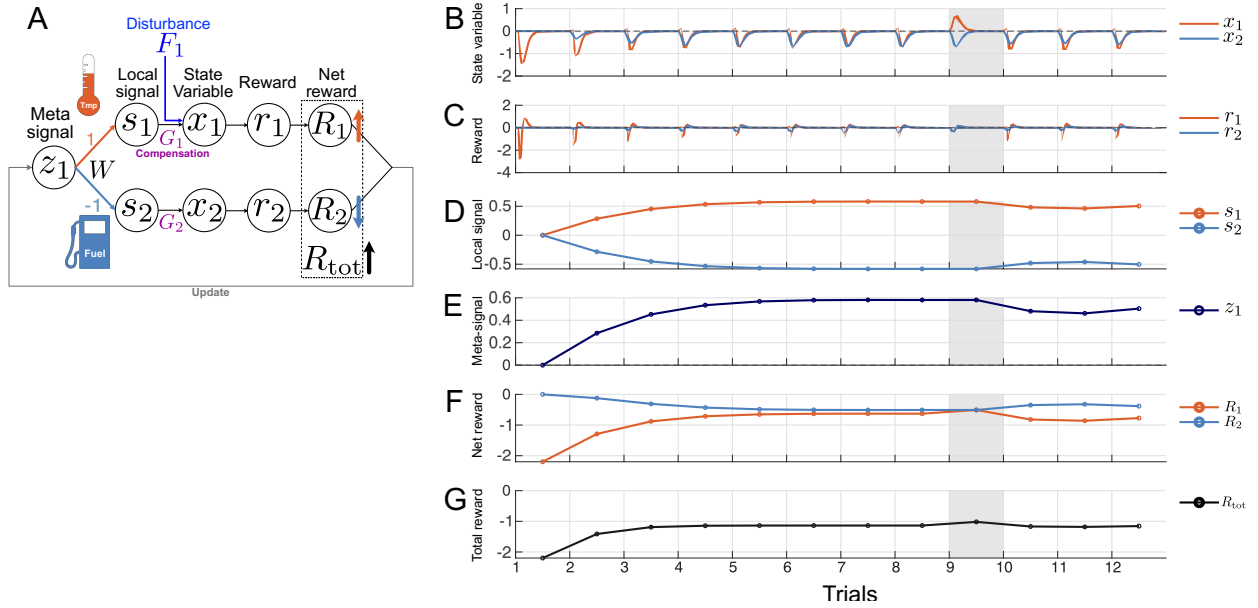


図 10. メタ信号による複数変数のホメオスタシス調節 (A) メタ信号 1 個・局所変数 2 個からなるモデル構成。シミュレーションで用いた共有制御の単純設定 ($K = 1, N = 2$) を示す。単一のメタ信号 z_1 が、符号の異なる影響を介して 2 つの局所変数を同時に調節する。低体温の例では、 z_1 の増加は体温を上昇させる一方で、抽象的なエネルギー関連資源を消費するものとして解釈される。(B–G) 報酬重み付けパラメータ $\eta = 1.5$ の下で得られたモデル変数のシミュレーション時系列。網掛け領域は消去期を示す。(B) 局所変数 x_n の時系列 (第 1 変数を橙、第 2 変数を青で表示)。(C) 瞬時報酬 $r_n(t)$ の時系列。(D) 局所信号強度 s_n の試行ごとの変化。(E) x_1, x_2 を調節するメタ信号 z_1 の推移。(F) 各局所変数に対する試行総報酬 R_n 。(G) 系全体の総報酬 R_{tot} 。

5.2 共有メタ信号制御下における優先度付けの効果

次に、共有制御を伴う多次元環境において、内部変数間の「優先度」が予期的調節にどのような影響を与えるかを検討した。第四章 4.4 の拡張設定、2つのメタ信号 (z_1, z_2) , $(K = 2)$ により3つの内部変数 (x_1, x_2, x_3) , $(N = 3)$ が制御される系を考えた。外乱は、これまでと同様に x_1 にのみ直接作用する。したがって、外乱を受ける変数の調節を改善しようとする試みは、他の変数を各自の基準値近傍に保つ要求と必然的に競合する。この競合構造を定量的に調べるため、各変数に割り当てる優先度重み a_n を系統的に変化させ、トレードオフ、回復ダイナミクス、および総報酬 R_{tot} がどのように変容するかを評価した。

まず、優先度が強く偏った場合にトレードオフ構造がどのように再編されるかを検討した。具体的には、非一様な優先度設定として、各変数を順に「最優先」とする条件をシミュレートした（図 11）。最初に、外乱の直接対象である x_1 を最優先とした場合、すなわち $a_1 > \max(a_2, a_3)$ のとき、高い優先度が割り当てられたドライブ関数 D_1 を含む反動的ドライブ項 $\partial D_1 / \partial x_1$ が強く働くことで x_1 の偏差は大きく抑制され、基準値への回復も迅速となった（図 11B）。しかしその一方で、補償に伴う偏差は x_2 および x_3 側へ“押し出される”形で増大し、総報酬 R_{tot} の増加は限定的であった。対応する制御信号およびメタ信号の試行ダイナミクスは図 12 に示した。

これに対し、 x_2 または x_3 を最優先とした場合、すなわち $a_2 > \max(a_1, a_3)$ または $a_3 > \max(a_1, a_2)$ のとき、系は最優先変数を保護する方向に調節を配分し、外乱を受ける x_1 の偏差はより大きく、回復も遅延した（図 11C, D）。その結果、報酬の蓄積は遅れ、低優先度とされた変数群における偏差も増大した（詳細は図 13, 14）。より詳細を見てみると、回復の遅延は、 $a_2 > \max(a_1, a_3)$ よりも $a_3 > \max(a_1, a_2)$ の方がさらに大きい。これは、外乱を受ける x_1 に対する調整の影響が、メタ信号を介して、近い変数から遠い変数へ伝播していくためと考えられる。

3つの内部変数 (x_1, x_2, x_3) の優先度が概ね均衡している場合についても確認した。系は質的な破綻を伴わず安定に調節された。外乱が直接作用するのは x_1 のみであるにもかかわらず、メタ信号を介した共有制御により x_2 および x_3 にも偏差が生じた。その結果、補償に伴うコストは複数変数へ分散し、各変数が中程度に基準値から逸脱しつつも、総報酬 R_{tot} は比較的高い水準に保たれた（図 15）。

以上より、共有メタ信号制御下では、優先度付けは補償コストを「消去」するのではなく、どの変数にコストが配分されるかを再配分することが示された。極端な優先度付けは高優先度変数の保護には有効である一方、外乱を受けた状態の有効な補償を阻害し得る。すなわち、「特定変数の保護」と「全体としての調節性能」との間に内在的なトレードオフが存在することが明確になった。

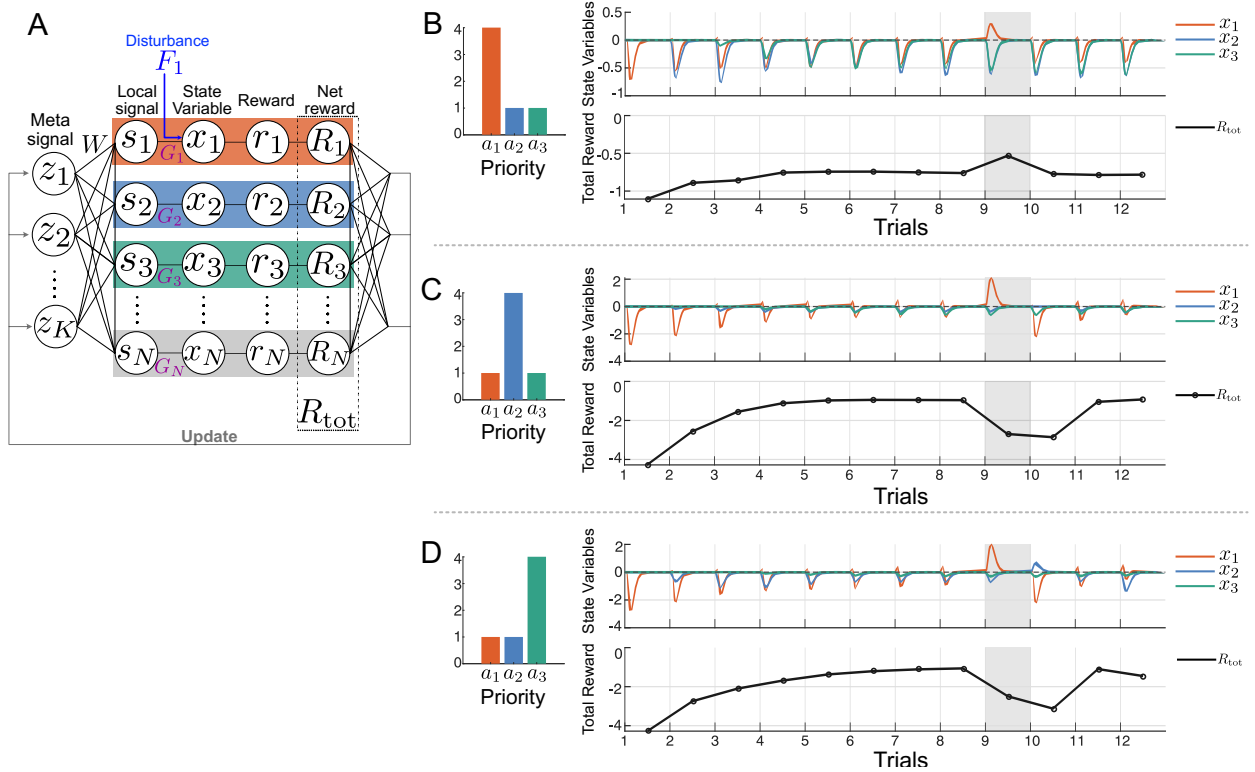


図 11. 3 種の優先度重みパターン下における複数変数のホメオスタシス調節 (A) K 個のメタ信号が N 個の局所変数を制御するモデル構成。シミュレーションでは $K = 2, N = 3$ とした。(B-D) 3 種の優先度設定におけるシミュレーション結果。棒グラフ (左) に示すように、変数 x_1 (B)、 x_2 (C)、または x_3 (D) をそれぞれ最優先 (最大重み) とした。赤・緑・青の棒はそれぞれ x_1, x_2, x_3 に対する優先度重みを表す。各設定について、右上パネルは各変数の基準値からの偏差 $x_n(t) - x_n^*$ の時系列を、右下パネルは試行にわたる総報酬 R_{tot} の推移を示す。網掛け領域は消去試行を示す。

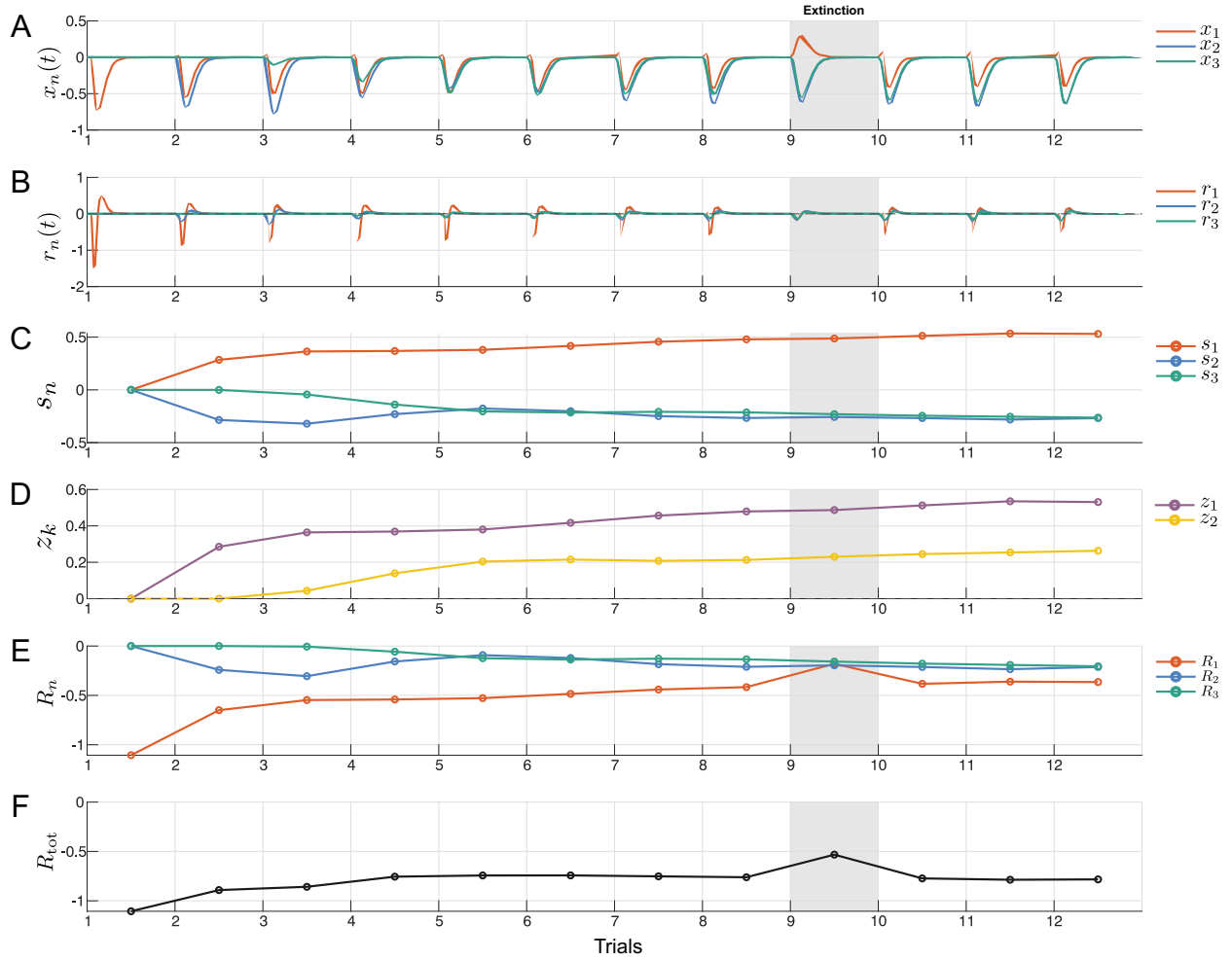


図 12. 多次元モデルの各変数に非一様な優先度を設定した場合のパラメータの挙動, $a = (4, 1, 1)$. (A) 試行を通じた内部変数 $x_n(t)$ ($n = 1, 2, 3$) の時系列。(B) 各変数について計算された瞬時報酬 $r_n(t)$ 。(C) 局所制御信号 s_n の試行ごとの推移。(D) メタ信号 z_1 および z_2 の試行ごとのダイナミクス。(E) 各試行において各変数から得られた純報酬 R_n 。(F) 試行を通じたシステム全体の総報酬 R_{tot} 。網掛け領域は消去試行を示す。本図は、図 11B に示した非一様優先度条件に対応する。

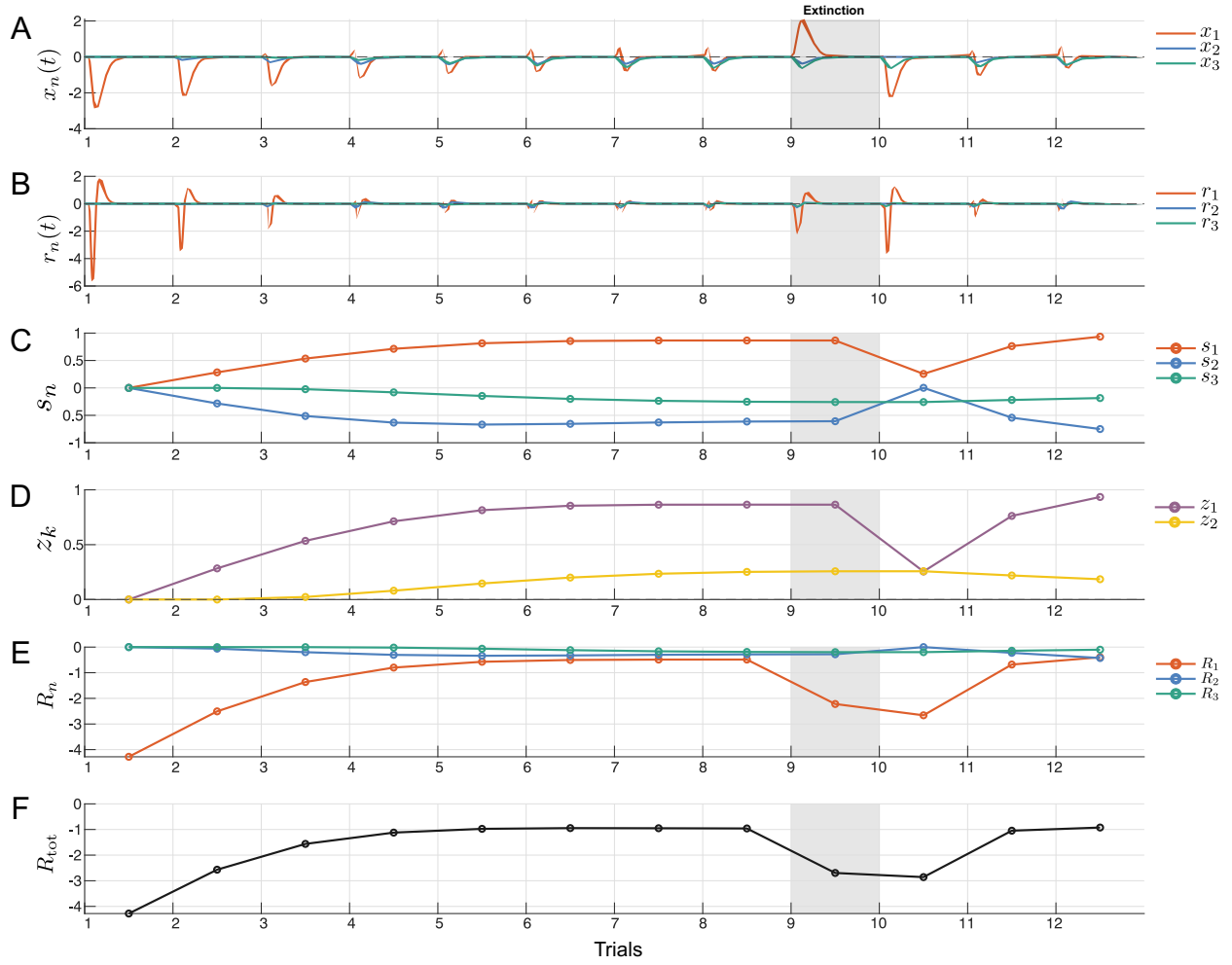


図 13. 多次元モデルの各変数に非一様な優先度を設定した場合のパラメータの挙動, $a = (1, 4, 1)$. (A) 試行を通じた内部変数 $x_n(t)$ ($n = 1, 2, 3$) の時系列。(B) 各変数について計算された瞬時報酬 $r_n(t)$ 。(C) 局所制御信号 s_n の試行ごとの推移。(D) メタ信号 z_1 および z_2 の試行ごとのダイナミクス。(E) 各試行において各変数から得られた純報酬 R_n 。(F) 試行を通じたシステム全体の総報酬 R_{tot} 。網掛け領域は消去試行を示す。本図は、図 11C に示した非一様優先度条件に対応する。

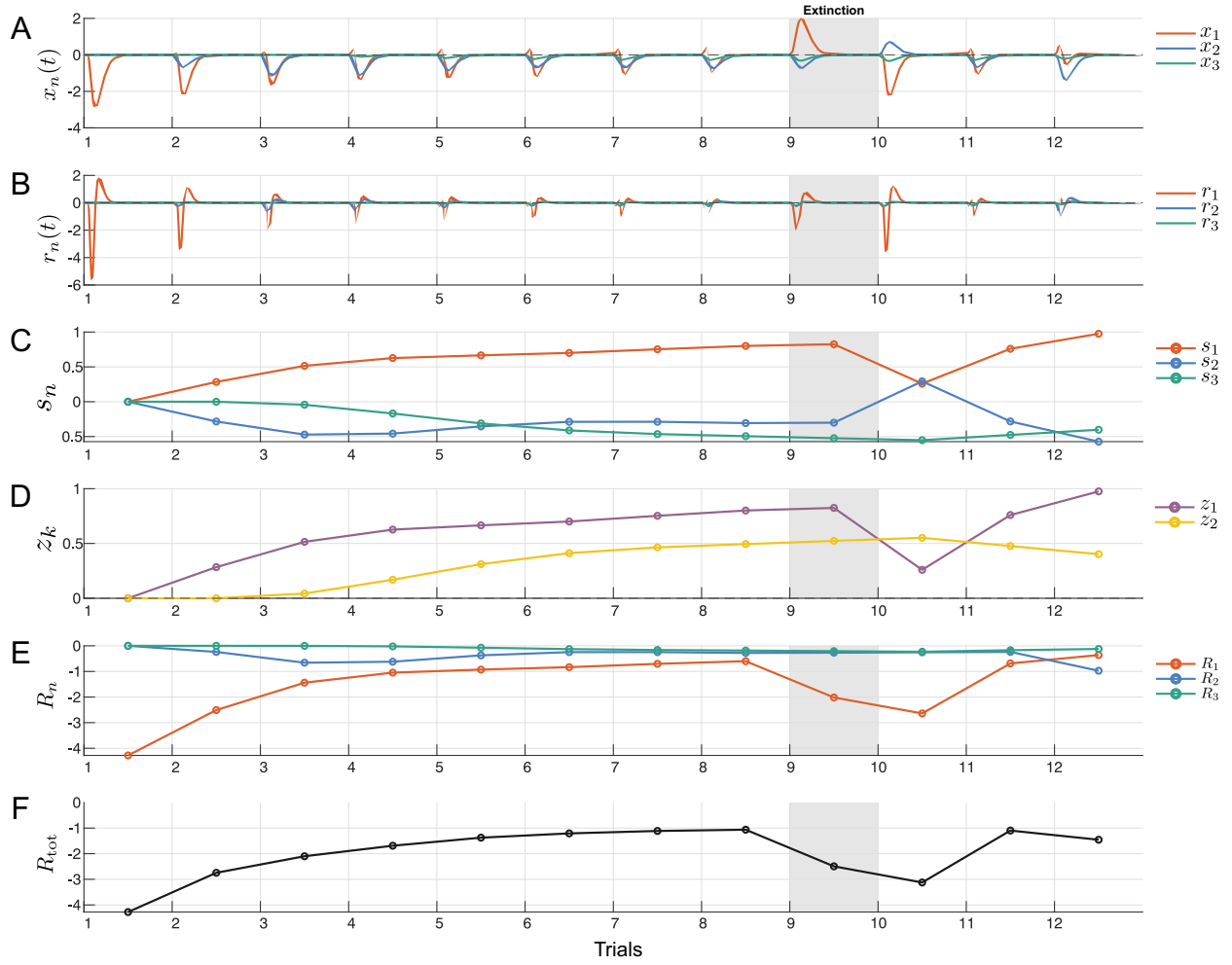


図 14. 多次元モデルの各変数に非一様な優先度を設定した場合のパラメータの挙動, $a = (1, 1, 4)$. (A) 試行を通じた内部変数 $x_n(t)$ ($n = 1, 2, 3$) の時系列。(B) 各変数について計算された瞬時報酬 $r_n(t)$ 。(C) 局所制御信号 s_n の試行ごとの推移。(D) メタ信号 z_1 および z_2 の試行ごとのダイナミクス。(E) 各試行において各変数から得られた純報酬 R_n 。(F) 試行を通じたシステム全体の総報酬 R_{tot} 。網掛け領域は消去試行を示す。本図は、図 11D に示した非一様優先度条件に対応する。

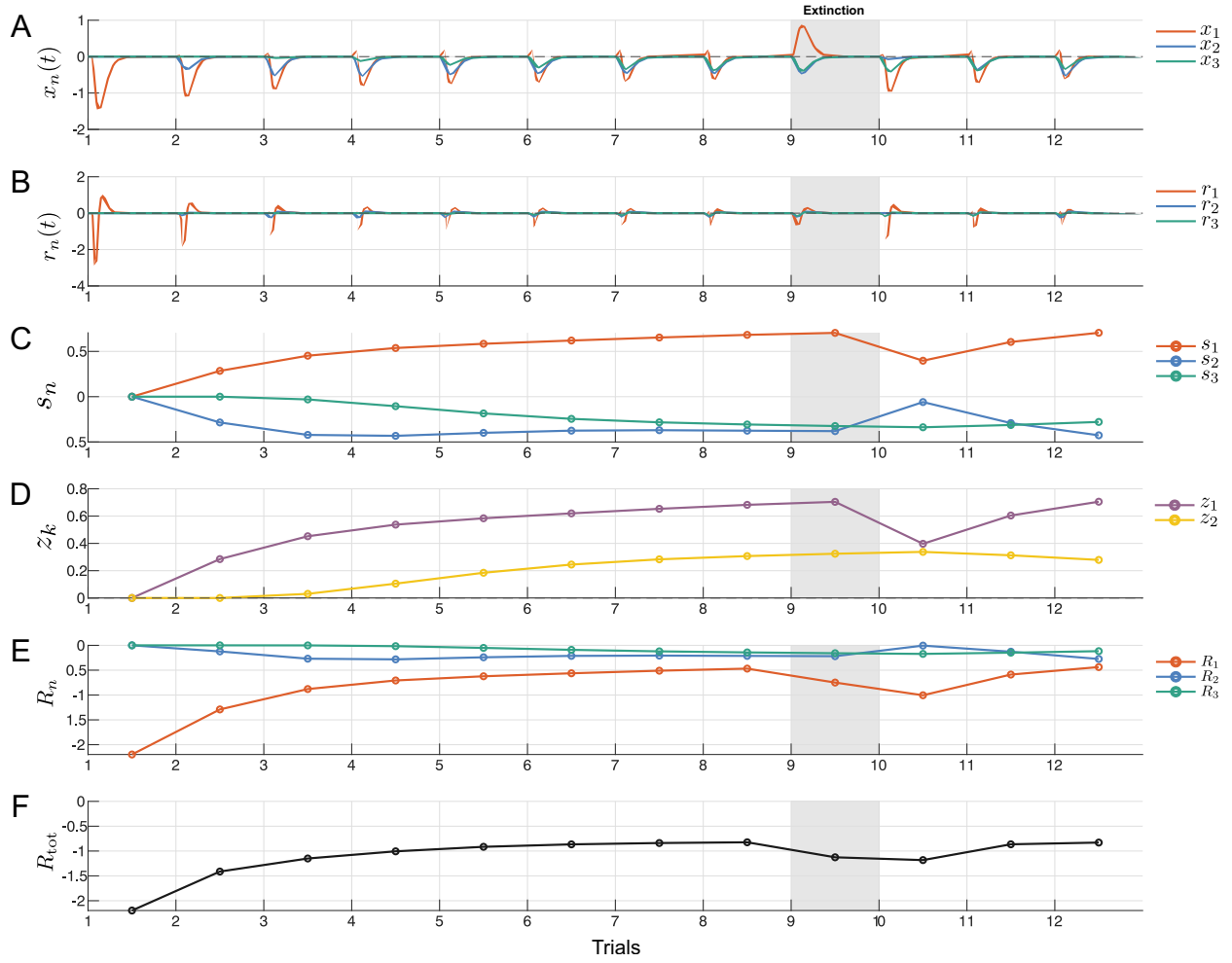


図 15. 多次元モデルの各変数に一様な優先度を設定した場合のパラメータの挙動, $a = (2, 2, 2)$. (A) 試行を通じた内部変数 $x_n(t)$ ($n = 1, 2, 3$) の時系列。(B) 各変数について計算された瞬時報酬 $r_n(t)$ 。(C) 局所制御信号 s_n の試行ごとの推移。(D) メタ信号 z_1 および z_2 の試行ごとのダイナミクス。(E) 各試行において各変数から得られた純報酬 R_n 。(F) 試行を通じたシステム全体の総報酬 R_{tot} 。網掛け領域は消去試行を示す。

5.3 共有制御下における学習の成功条件と失敗条件

強化学習が、多次元にわたる予測的調節をどのような条件で適切に協調・獲得できるかを明らかにするため、報酬非対称性 η と優先度角 θ の同時効果を系統的に検討した。各パラメータ組 (η, θ) について、まずメタ信号空間における総報酬のランドスケープを特徴づけ、報酬を最大化する配置 $\mathbf{z}^*(\eta, \theta)$ を同定した (図 16B)。探索した η の範囲全体において、最適点 $\mathbf{z}^*(\eta, \theta)$ の位置は主として優先度パターン θ に依存し、 η を変化させても大部分が重なった。そこで、報酬最適点の幾何学的構造を示すため、 $\eta \in [1, 3]$ を一様離散化した 15 点のうち 5 番目の値として定義した代表値 $\hat{\eta} = 1.5714$ に注目して解析を進めた。

まず、 $\eta = \hat{\eta}$ とし、 (z_1, z_2) の 100 試行の軌跡を追跡した。その結果、ほとんどの優先度角 θ において初期値 $(z_1, z_2) = (0, 0)$ から 報酬最適配置 $\mathbf{z}^*(\theta | \eta = \hat{\eta})$ に向かって軌跡が発展することが確認された (図 17B)。一方、優先度が強く偏った一部の θ では、 $\mathbf{z}^*(\theta | \eta = \hat{\eta})$ に接近しつつも最適点を通過する挙動が見られた (図 17B, 右から 2, 3 番目)。続いて、学習が 100 試行で収束しているかを評価するため、 (z_1, z_2) の一次差分ユークリッド距離を計算した。距離は時折スパイクを示したものの、全体として増大傾向はなく、最終的に十分に小さな値 (概ね < 0.01 程度) へ収束した。例外的に、100 試行あたりで大きな跳躍を示す θ (図 17C, 右から 5 番目) が存在したが、追加で 100 試行 (計 200 試行) をシミュレーションしたところ、その後収束が確認された。以上より、本系は少なくとも本観測範囲において、顕著な振動や発散を示さず、収束として学習が完了することが示された。このことは、各試行における (z_1, z_2) と報酬最適配置 $\mathbf{z}^*(\theta | \eta = \hat{\eta})$ との距離の推移 (図 17D) とも整合的である。

次に、報酬非対称性 η と優先度角 θ の同時効果を俯瞰視するため、報酬最適配置 $\mathbf{z}^*(\theta | \eta = \hat{\eta})$ と、最終的な収束値を図示した (図 18A, B)。優先度が概ね均衡している領域 (図 18A) と優先度が強く偏った領域 (図 18B) とを分けてある。優先度が概ね均衡している領域では、報酬最適配置 $\mathbf{z}^*(\theta | \eta = \hat{\eta})$ は $z_1 - z_2$ 平面上で滑らかな楕円状の多様体を形成した (図 18A)。この条件では、学習によって得られるメタ信号の軌跡が対応する最適点付近へ安定に収束し、競合する内部変数間のトレードオフを適切に折衝できることが示された。これに対し、優先度が強く偏った領域では、報酬最適点自体はメタ信号空間内で概ね優先度が均衡している場合と同様の位置に存在するにもかかわらず、学習軌跡はそれらの最適点から大きく逸脱した (図 18B)。この乖離は、優先度が過度に偏ると、強化学習が補償調節を協調させることに失敗することを反映している。

最後に、全パラメータ空間にわたって学習性能を定量化するため、各 (η, θ) において、報酬最適配置 $\mathbf{z}^*(\eta, \theta)$ と学習の最終収束点との間のユークリッド距離の平均を指標として用いた。得られた相図 (図 18C) では、 (η, θ) 平面上において、提案する強化学習機構が一貫して報酬最適配置に近い解へ到達する領域が連続的に現れ、これは優先度が均衡している、あるいは中程度に不均衡な条件に対応した。これに対し、極端な優先度付けを特徴とする領域では、最適点への収束が系統的に失敗し、軌跡は安定ではあるものの最適から外れた配置へ落ち着いた。以上により、多次元系において予測的ホメオスタシス調節が学習可能となるパラメータ領域が明確に画定された。

総括すると、共有制御下では優先度付けが学習成立性を決定づける主要因となる。すなわち、報酬最適配置そのものは広いパラメータ範囲で存在する一方、優先度が極端に偏ると、学習がその近傍へ到達できなくなる。このことは、定常状態の解析だけでは捉えにくい、予測的調節の失敗モードが存在することが示唆された。

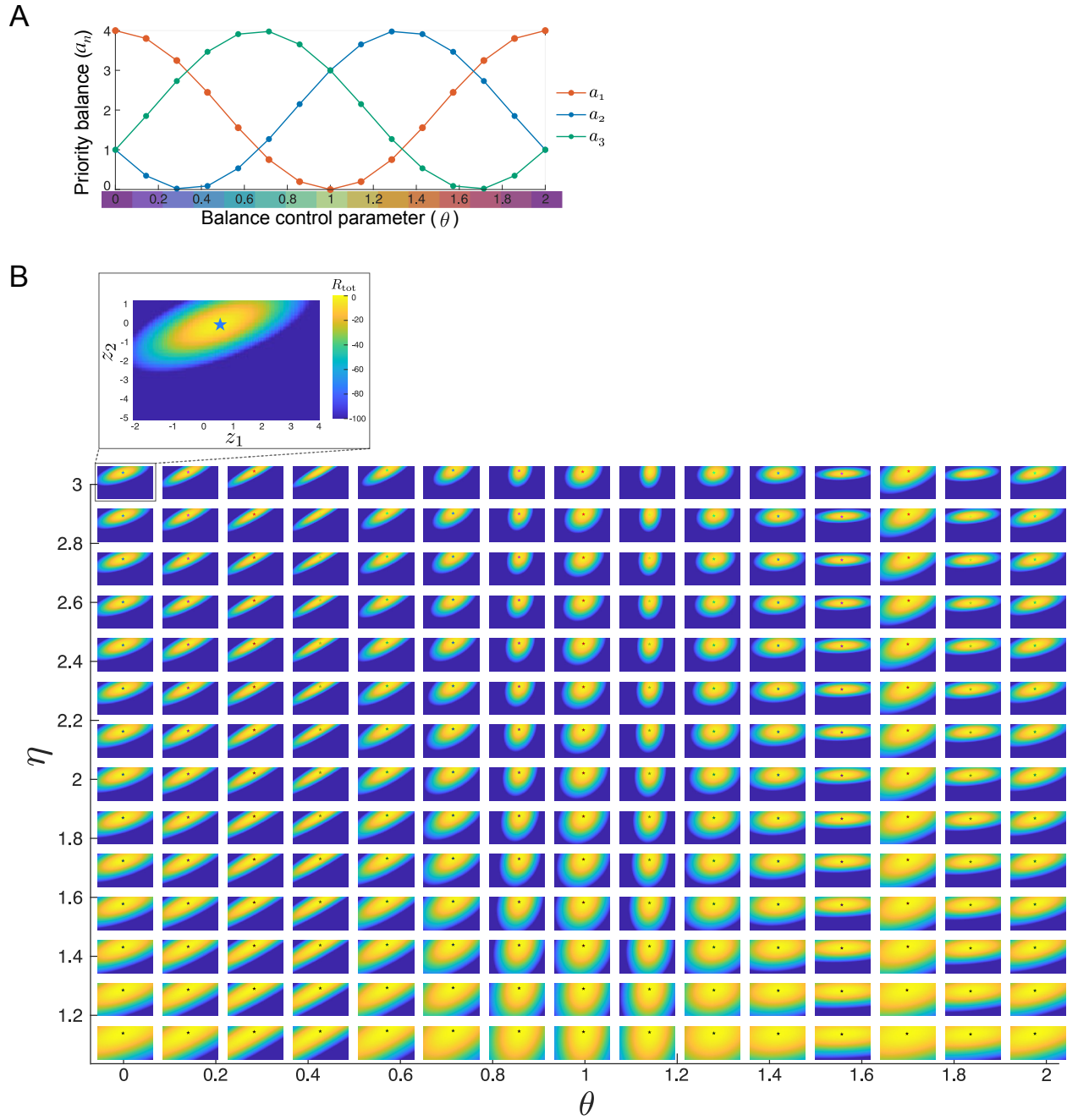


図 16. 優先度付けと報酬非対称性の系統的解析のための準備 (A) 優先度重み a_n の、優先度角 θ によるパラメータ化。橙、緑、青の曲線はそれぞれ a_1, a_2, a_3 を表す。(B) メタ信号空間における報酬ランドスケープ代表的な (η, θ) のパラメータペアについて、総報酬 R_{tot} を (z_1, z_2) の関数として示したヒートマップ。これらのマップを用いて、図 12 に示す報酬最適メタ信号配置を同定した。 $\eta = 1$ は、一次元モデルの結果により学習が成立しないことが示されたため除外した。

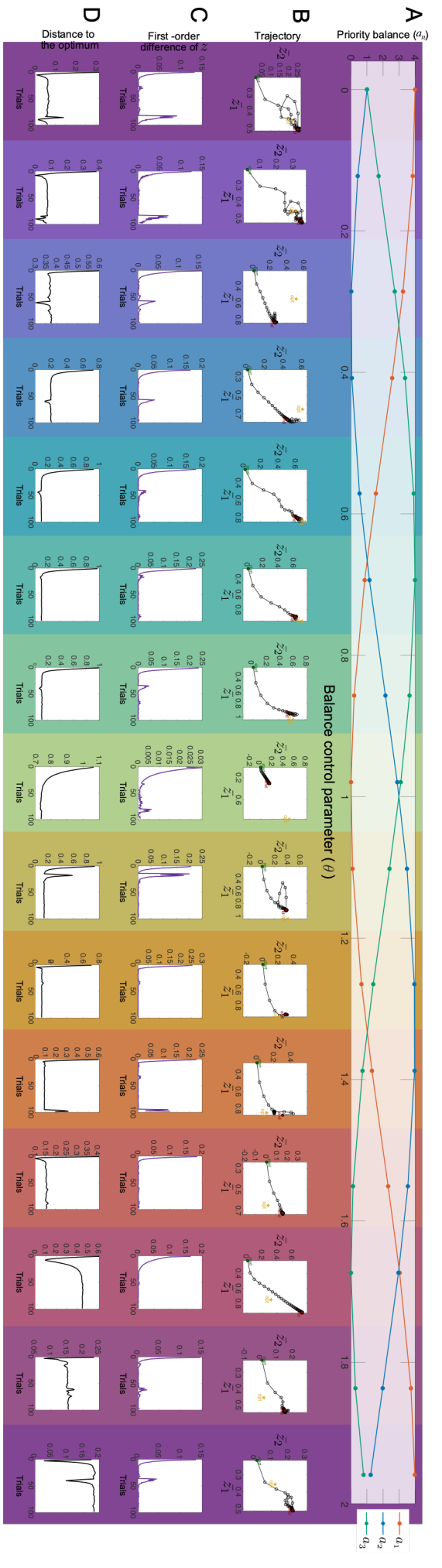


図 17. $\eta = \hat{\eta}$ のときの全 θ の軌跡. (A) 優先度角 θ と対応する優先度重み a_n 。背景色は、優先度角 θ 15 サンプルそれぞれに図 16, 18 で用いたものと共通の配色を用いた。
 (B) メタ信号の重心の軌跡。シミュレーションの開始時点 (緑円)、終了時点 (赤丸)、 $z^*(\theta, \hat{\eta})$ (黄星)。(C) メタ信号のユーリッド距離の一次差分。(D) 報酬最適配置 $z^*(\theta, \hat{\eta})$ との距離。

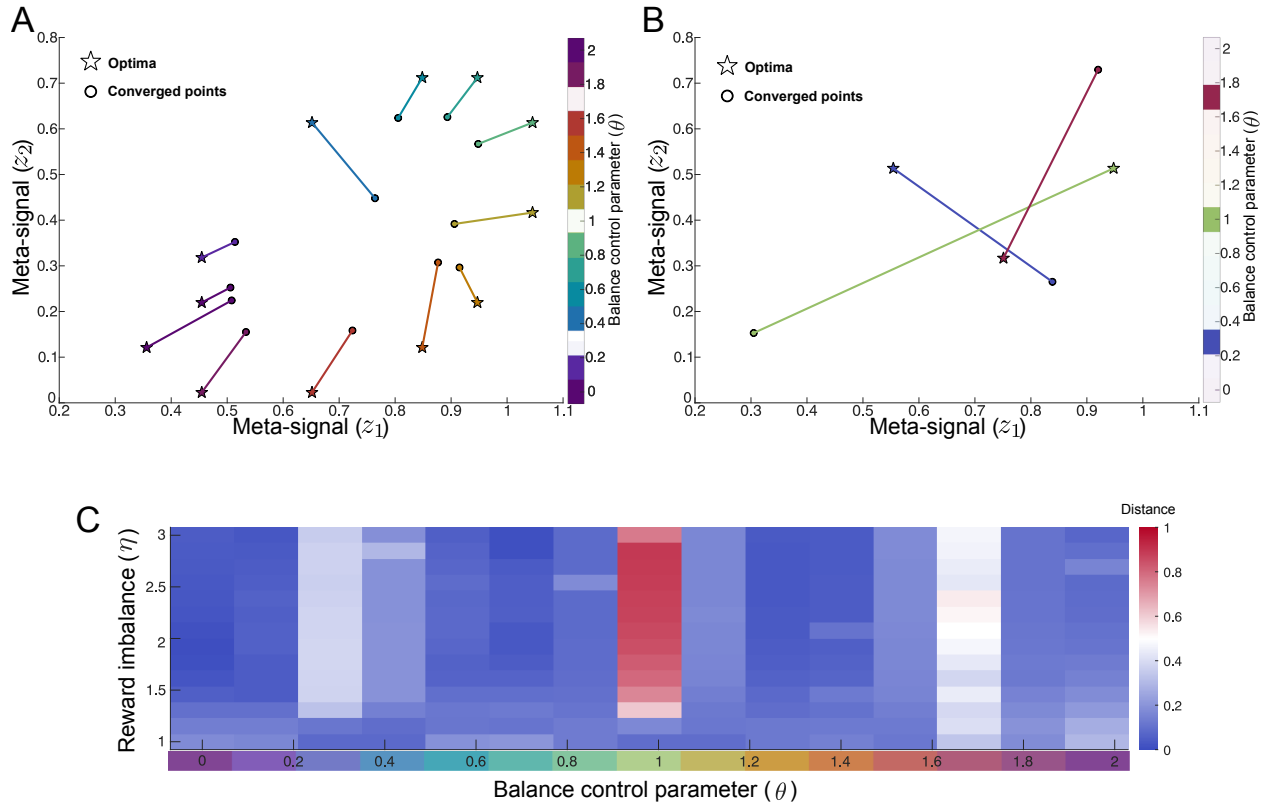


図 18. パラメータ対が多次元モデルのダイナミクスに及ぼす影響 (A, B) $\eta = \hat{\eta}$ を固定したときの報酬最適点 $z^*(\theta | \eta = \hat{\eta})$ (星) と、学習後に収束した点 (円) を示す。両者を結ぶ実線は収束点と最適点までの距離を表す。欠損色は、図中に表示していない θ を示す。(C) (η, θ) 平面上における、収束点と報酬最適点 $z^*(\eta, \theta)$ との平均距離のヒートマップ (最終 10 試行の平均)。

第六章

考察

6.1. 主要な知見と新規性

本研究は、経験に基づく予期的かつ漸進的な内部生理状態の調整を可能にする、拡張版の恒常性強化学習（homeostatic reinforcement learning, HRL）の枠組みを提案した。一次元の体温調節条件づけ課題において、本モデルは、試行ごとに補償応答を学習することにより、エタノール耐性の獲得および消去後の迅速な再獲得を再現し、CS-US に基づく体温調節データ (Mansfield & Cunningham, 1980) と整合した。さらに複数の生理軸へ拡張すると、共有された調節信号は不可避なトレードオフを課することが示された。すなわち、優先度構造は、どの変数が厳密に安定化され、どの変数が変動を許容されるかを規定し、その結果として学習の収束あるいは失敗の様式を決定した。

6.2. 結果の解釈

シミュレーション結果は、提案モデルが経験依存学習による適応的生理調節の本質的特徴を捉えることを示した。一次元設定では、モデルはエタノール誘発低体温に対する耐性の漸進的な形成を再現し、予期される攪乱に先行して代償反応を能動的に活性化することを学習した。重要なのは、この連続かつ漸進的な代償信号の調整が、従来の離散 HRL では説明困難だった耐性獲得の段階性を捉えられる点である。したがって、本モデルは、先行的応答が経験を通じて獲得される過程について、より生物学的に妥当な説明を与える。

さらに、本定式化では、ドライブ増加（悪化）に対して同程度のドライブ減少（改善）よりも大きな重みを置く、報酬の非対称評価 ($\eta > 1$) は学習において決定的な役割を果たした。この非対称性がない場合、時間微分に基づく報酬が情報量を失い、学習が成立しない。一方、この仕組みがある場合には、エージェントは大きな逸脱を予防する方策を学習し、生物で観察されるリスク回避的行動に呼応する振る舞いを示した。

また、活性化・抑制の二つのユニットによる消去学習に関しては、文脈感受的なゲーティングが鍵となった。ゲートを導入することで消去学習と迅速な再獲得を同時に説明できた一方、両ユニットが存在してもゲートがない場合には迅速な再獲得は生じなかった。これは、抑制トレースの形成と、その表出の切替えを分離して扱う必要性を示している。本研究では解析的明瞭性のため、ゲートを二値の ON/OFF として導入したが、表出の切替えがスイッチ様に観察され得ることは、連続的に変化する抑制強度や回路状態が閾値や競合といった非線形性を介して急峻に反映されることで説明できる。したがって、本モデルの二値ゲートは表出切替えの近似と位置づけられる。将来的には、ゲートを連続量として学習させ、非線形変換により二値的表出が生じる拡張が考えられる。なお、これらの二つのユニットとゲーティング機構が、文脈依存的な恒常性調節へどこまで一般化可能かは今後の課題であるが、本研究は、少なくともパブプロフ型の身体状態調節において、形成と表出の分離を含む類似の枠組みが必要であることを示唆する。

多次元拡張では、共有制御信号（メタ信号）が内部変数間のトレードオフを生む機序を示した。ある恒常性変数に高い優先度重みを付与すると、系はその変数の誤差を強く抑制するよう学習する一方で、相対的に他の変数の逸脱を許容しやすくなる。これは、体温や血圧のような重要変数を厳密に維持する代わりに、重要度の低い変数の変動を許すという、生物学的に妥当な戦略に合致する。しかし、特定の変数の優先度が極端に低い場合には、その誤差がほとんどコストとして評価されず、全体として最適値への収束が阻害されうる。以上の力

学は、優先度バランスが過度に偏ると適応が失敗する理由を説明し、各変数に最低限の規制重みを維持する重要性を示唆する。

さらに、特定の恒常性変数の優先度が持続的に低い場合、その逸脱は十分に是正されず、モデル上は当該誤差が「介入しても減らせない」ものとして扱われるため、系はそれを不可制御あるいは外因性として推定しやすい。その結果、環境が不安定であるという推定が固定化され（シミュレーションでは環境変化確率 ω が高止まりした）、内部状態と有効な代償反応の連合形成に失敗することを示した。逆に、誤差に過大な重みが割り当てられる場合には、外因性の変動までも内因化して過剰な代償反応を持続させ、調節出力の過大化や硬直化を通じて別種の適応失敗（過剰制御）を招き得る。これらの含意を臨床例と照合する議論は次節で行う。

6.3. 病理との関連

前節で示したように、多次元モデルでは優先度バランスの偏りが外因/内因の帰属と予期的調節の収束性に影響し得る。特に、ある恒常性変数に対して不釣り合いに低い実効的優先度が割り当てられると、その偏差は学習過程で過小評価され、共有かつ低次元 ($K < N$) の制御の下で擾乱の原因帰属が誤りやすくなり、予測的調節が収束しない（非収束的適応）状況が生じ得る。

筋痛性脳脊髄炎/慢性疲労症候群（myalgic encephalomyelitis/chronic fatigue syndrome: ME/CFS）は、この種のダイナミクスと接点を持ち得る臨床例として位置づけられる。臨床的に ME/CFS は、起立性不耐やその他の自律神経系の異常と関連し (Freeman & Komaroff, 1997; Garner & Baraniuk, 2019)、また疲労の重症度と関連する心拍変動（heart-rate variability）の低下も報告されている (Escorihuela et al., 2020; Nelson et al., 2019)。より広い文脈では、近年の見解は、単純な感覚障害というよりも、アロスタシス調節の障害や内受容推論の異常に重点を置いている (Annesley et al., 2024; Arroll, 2013; Barrett & Simmons, 2015; Goertzel et al., 2006; Sharp et al., 2021; Stephan et al., 2016)。

重要なのは、本モデルが特定の変数の生理学的重要性が文字通り低下することを含意するわけではない点である。むしろ本研究では、そのような病態を計算論的に、内受容あるいは自律神経制御の変調に起因して学習過程における「実効的な優先度」が不均衡となる状態として解釈している。この枠組みによれば、臨床症状をモデルの特定のパラメータ領域に対応づけることが可能となり、観察される自律神経失調を固定的な感覚欠損ではなく、基底にある学習ダイナミクスと結びつけて理解できる。原理的には、摂動を伴う学習パラダイムと生理時系列データを組み合わせることで、非収束的適応が特定の患者集団に特徴的か否かを評価し得ると考えられる。

6.4. 先行研究との比較

恒常性維持の古典的モデルは、主として制御理論 (Negative feedback) に基づき、内部セットポイントからの逸脱に対する反応的調節を記述してきた (Ashby, 1960; Cannon, 1932; Guyton et al., 1973)。これらは、例えば圧受容体反射など即時的な生理調節の説明に成功してきたが、経験や学習、文脈適応を取り込む機構を一般に欠いており、薬物耐性の漸進性や文脈依存的調節を十分に説明できない。

この限界に対し、アロスタシスが恒常性理論の拡張として提案された。恒常性が「不変による安定」を強調するのに対し、アロスタシスは「変化による安定」を強調し、環境需要を予測した文脈依存的な設定点調整の必要性を唱える (McEwen, 2017; McEwen & Wingfield, 2003; Ramsay & Woods, 2014; Sterling, 2004; Sterling & Eyer, 1988)。この枠組みは、ストレス生理学や神経内分泌系における予期的調節を支持する実証研究に裏付けられている。さらに、Ganzel ら (2010) は神経科学・心理学・社会科学のモデルを統合し、アロスタシスがヒト脳のストレス応答を理解するための統一原理となると主張した (Ganzel et al., 2010)。もっとも、アロスタシスは主として概念的・記述的な枠組みに留まり、反復経験から予期的調節がどのように学習されるかについての厳密な計算論的実装は乏しい。本研究で提案するモデルは、固定のセットポイントを持つ設定であるため、一見、適宜セットポイントを変動させるようなアロスタシスの考え方に反するよう見えるかもしれない。しかしながら、多次元拡張原理を強化学習により具現化し、経験依存更新から先行的調節が自然に立ち現れることを示す点で、このギャップを補う。

近年の予測符号化や自由エネルギー原理 (FEP) は、世界と身体の内部生成モデルを維持・更新して予測誤差を最小化する推論システムとしてエージェントを捉え、先行的制御への別経路を提供する (K. J. Friston & Stephan, 2007; K. Friston et al., 2015; Pezzulo et al., 2024; Tschantz et al., 2022)。これらは規範的かつ生物学的動機付けの強い説明力を持つ一方、事前構造化された/学習された潜在動態モデルを仮定し、計算コストの高い推論を要することが多く、自律・代謝といった迅速で資源制約の厳しい調節系に適用するには難がある。

対照的に、本研究の強化学習ベースのアプローチは、明示的な生成モデルを構築せずに先行的制御を達成する。具体的には、スカラーのドライブ関数と観測結果から直接得られる強化信号に依拠し、単純なヒューリスティック (ω) によって報酬変化の帰属 (内部調節か外的擾乱か) を区別する。にもかかわらず、この仕組みは、耐性の漸進的獲得や文脈依存的な柔軟調節を捉えつつ、計算量は小さくかつ生物学的妥当性を保持する。すなわち、本枠組みは、予測符号化/FEP に対するモデルフリーな補完として機能し、後者が統一的・強力な理論的説明力を持つ一方で生成モデルに依存するのに対し、我々はより単純なモデルフリー強化学習からも先行的特徴が立ち現れることを示す。

6.5. 限界と今後の課題

本モデルは先行的恒常性調節のいくつかの重要側面を捉える一方、以下の限界がある。それらについて、今後の課題と展開を示す。

第一に、報酬をドライブ関数の時間微分 (dD/dt) のみに基づいて定義しているため、定常的な逸脱に対する感度が低く、同時に、理想的な状態に留まっていることが評価されない。

将来的には、 D と dD/dt の双方を併用することで、持続的逸脱と短期的変化を区別する拡張が考えられる。

第二に、メタ信号の影響行列 $\mathbf{W}$ が固定である点である。これは、共通のメタ信号が改善させる局所変数と、そのために犠牲となる局所変数を結ぶ結合を表現し、多変数のトレードオフを通じて有効な非単調調節を生じさせる役割がある。より具体的には、 $\mathbf{W}$ はメタ信号が局所変数へ及ぼす固定の生理学的効力を表すのだが、固定であるために、効果器側そのものの変化（例：自律神経出力の低下）を表現しない。すなわち、季節適応のような、より遅い文脈変化や環境変化、個体差などは $\mathbf{W}$ として表現できないが、それは優先度重み a の変動として表現する方が自然である可能性がある。

第三に、本研究はシミュレーション検証に留まり、多変数の実生理データへの当てはめは行っていない。ただし、モデル構造はそのようなデータとの整合性が高く、定量的適合に拡張可能である。今回の研究で、報酬非対称性 η と 優先度角 θ の組み合わせによっては、学習を経ても最適値に到達しないという現象が生じた。これらのパラメータ領域は、実際の生理系における不適応・調節破綻に対応する可能性がある。慢性ストレス、代謝異常、神経変性疾患等における恒常性破綻の実データと照合することが、重要な次の一歩である。加えて、観察されるトレードオフや優先度戦略が再現されるか、また、固定の生理学的効力と適応的優先度付けの区別を検証することは将来課題である。

第四に、本研究では、局所変数そのものを能動的に回復させる（回復ゲインを上げる、あるいは回復資源を追加投入する）機構をモデルに組み込まなかった。その結果、擾乱が反復される環境では、不可避なトレードオフが系全体に分配される一方で、条件によっては特定の局所変数が継続的に負担を引き受ける、いわば「痛み分け」の偏りが生じ得る設定となった。このような偏りは、現実の生体においては、脆弱な臓器における損耗や、臨床的に観察されやすい症状として表出する負荷の集中として解釈し得る（アロスタシス・アロスタティック負荷の枠組み）。一方で、生体は反復負荷の下でも長期的な破綻を回避する必要がある、特定の局所変数が恒常的に犠牲を強いられ続けられないための“回復の窓”が存在すると考えるのが自然である。概日周期や睡眠-覚醒周期は、代謝・免疫・修復関連過程を時刻に応じて切り替える「時間分割」として機能し、睡眠時に代謝産物クリアランスが促進されること等から、回復過程を支える候補機構となる(Raizen et al., 2023)。生体システムは、調節対象の変数数 N に対して制御信号（メタ信号）数 K が小さい $K < N$ という意味で、共有資源に基づく低次元制御を強いられる。この状況では、各変数の同時最適化は一般に保証されず、残差的偏差やトレードオフはアルゴリズムの失敗というより構造的制約の帰結である。そこで本研究の枠組みを拡張する方向として、概日・季節といったリズムにより優先度や回復ゲインが時間変動され、ある時間帯では特定変数を重点的に回復させ、別の時間帯では他の変数へ資源配分を切り替えるという時間共有的な調節を導入すれば、負担の固定化を避けつつ全体性能を維持するという仮説が得られる。この観点から、本枠組みの将来拡張として少なくとも二つの方向が自然に導かれる。まず、優先度重みを定数 a_n ではなく、概日・季節リズムや文脈により緩やかに変動する $a_n(t)$ として導入することである。これにより、同一の共有制御信号の下でも、時間帯に応じて「どの偏差を強く罰するか／許容するか」が切り替わり、結果としてトレードオフの配分が時間平均的に平準化され得る。さらに、優先度重み $a_n(t)$ も制御対象として組み込み、強化学習を通じて $a_n(t)$ にリズムが創発するか、創発するならばどのようなリズムかを検証する。これによって、共有制御下のトレードオフを「不可避なコスト」として記述する段階から、時間構造を介した緩和・最適化の可能性へと議論を拡張できるだろう。

6.6. 生理学・モデリングへの含意

本モデルは、経験依存学習と生物学的制約を統合した先行的調節の計算ツールを提供する。適用領域として、体温調節、ホルモン性フィードバック、体液バランス、慢性的薬理効果など多岐にわたる可能性がある。強化学習と生理調節を橋渡しすることで、計算量の扱いやすさと生物学的基盤を兼ね備えた理論枠組みを提示する。これは、制御理論的・ベイズ的・予測符号化的アプローチを補完し、静的最適化や推論だけでなく、学習に基づく適応の重要性を強調するものである。また、本研究はシミュレーションに基づくが、本枠組みはモデル-データ統合への道を開く。実データへの適合によって説明力と予測力を検証・拡張することで、病態における調節不全の理解や介入戦略の設計に寄与しうる。したがって、本枠組みは、先行的恒常性調節の理論的理解を前進させるだけでなく、将来的なトランスレーショナル研究において、計算モデルと実生理の統合を促す基盤ともなりうる。

補遺

A1 一次元モデル

A1.1 漸進的予期的調節モデルの実装（提案モデル）

本付録では、第二章で導入した一次元の恒常性強化学習モデルについて、数学的定義を示す。再現性と解析的検討を担保するため、状態ダイナミクス、報酬定義、学習則、および数値実装の詳細をここにまとめる。

状態ダイナミクス

i 回目試行における体温状態 x のダイナミクスは次式で与える。

$$\frac{dx}{dt} = -\frac{dD}{dx} + f(t - t_{US,i}) + s_i g(t - t_{CS,i}).$$

関数 $f(t)$ および $g(t)$ は、 $t = 0$ を起点とする一過性関数であり、 $t \leq 0$ では $f(t) = 0$, $g(t) = 0$ する。 t_{US} および t_{CS} はそれぞれ US と CS の提示時刻を表す。 $f(t)$ はエタノール投与が体温へ及ぼす外乱（低体温）効果を、 $g(t)$ は補償反応を表す。外乱信号 $f(t)$ と補償反応 $g(t)$ は、共通の減衰率 k をもつ一次遅れ線形系のカスケードとして生成する。具体的には、観測された体温の波形を再現するため、低体温外乱 $f(t)$ は US 入力で駆動される 3 次カスケードの出力として定義する：

$$\begin{aligned}\dot{f}_1(t) &= -kf_1(t) + \xi_{US}(t), \\ \dot{f}_2(t) &= -kf_2(t) + f_1(t), \\ \dot{f}_3(t) &= -kf_3(t) + f_2(t), \\ f(t) &= f_3(t).\end{aligned}$$

同様に、予期的補償反応 $g(t)$ は CS 入力で駆動される 4 次カスケードの出力として定義する：

$$\begin{aligned}\dot{g}_1(t) &= -kg_1(t) + \xi_{CS}(t), \\ \dot{g}_2(t) &= -kg_2(t) + g_1(t), \\ \dot{g}_3(t) &= -kg_3(t) + g_2(t), \\ \dot{g}_4(t) &= -kg_4(t) + g_3(t), \\ g(t) &= g_4(t).\end{aligned}$$

ここで、 $\xi_{US}(t)$, $\xi_{CS}(t)$ はそれぞれ US, CS の入力信号を表す。これは、US, CS が提示されている時間窓の間は 1、それ以外は 0 となる二値パルスである。本研究では、各試行について上記の体温ダイナミクスを計算し、これを複数回繰り返すことで学習過程をシミュレートした。

ドライブ関数と報酬定義

ドライブ関数 $D(x(t))$ は、現在の状態が望ましい基準値 x^* からどれだけ離れているかに応じて身体状態を変化させようとする「動機づけ」を表す。本研究では二次形式を採用し、

$$D(x(t)) = a(x(t) - x^*)^2$$

と定義する。内部状態はこの基準値からの偏差として表し、正負はそれぞれ基準値より高い・低い側への逸脱を意味する。

モデルは制御変数（予期制御）を調整して報酬を最大化する。ホメオスタシス強化学習の枠組みに従い、時刻 t における瞬時報酬 $r(t)$ をドライブ関数の時間微分の負として定義する。さらに、プロスペクト理論に着想を得て、改善と悪化の評価の非対称性を導入し、罰（悪化側）に重みを付与する。瞬時報酬は

$$r(t) = \begin{cases} -\frac{dD}{dt} & \text{if } \frac{dD}{dt} \leq 0, \\ -\eta \frac{dD}{dt} & \text{if } \frac{dD}{dt} > 0. \end{cases}$$

とする。ここで $\eta > 1$ は非対称性を表すパラメータである。

i 回目試行の総報酬 R_i と、体温偏差の累積量 X_i は、それぞれの時系列を時間積分して求める：

$$R_i = \int_{t_{CS,i}}^{T_i} r(t) dt, \quad X_i = \int_{t_{CS,i}}^{T_i} (x(t) - x^*(t)) dt,$$

ここで T_i は i 回目試行の終了時刻を表す。本研究では固定時間窓 T_{trial} を用いる（具体的な値は表. 1 に記載）。

予期信号・抑制・ゲーティング

予期制御信号 s_i は試行レベルのスカラーであり、補償応答 $g(t)$ の振幅を調節する。 s_i は強化学習により試行間で更新される。補償信号の強度は、CS により駆動される 2 つの神経集団 A と I （それぞれ促進と抑制）によって調節されると仮定し、文脈依存のゲーティング変数 χ_i により相互作用が制御される。具体的には

$$s_i = A_i - \chi_i w^{A,I} I_i,$$

と定義する。ここで $w^{A,I}$ は I が A に及ぼす抑制効果の強さを表す。ゲーティング変数 χ_i は直前試行の CS-US 随伴性に基づき、

$$\chi_i = 1 - \text{CS}_{i-1} \text{US}_{i-1}, \quad \text{CS}_{i-1} \in \{0,1\}, \text{US}_{i-1} \in \{0,1\},$$

として定義する（式中の二値変数は、当該試行で CS・US が提示されたか否かを表す）。この機構は、「興奮性記憶が抑制性トレースによってマスクされ得る」こと、さらに「手がか

りの提示により抑制が一過的に解除され得る」ことを反映する。消去が元の連合を消去するのではなく、文脈依存的抑制を上書きするという生理学的知見を反映している。A および I は CS 入力により増強され、

$$A_i = u_i^A \text{CS}, \quad I_i = u_i^I \text{CS},$$

とする。\$u_i^A, u_i^I\$ はそれぞれ A と I の重みであり、Rescorla–Wagner モデルに代表される試行レベルの連合学習仮定に基づき更新される。重み \$u_i^{(m)}\$, \$(m \in \{A, I\})\$ の更新は

$$u_i^{(m)} \leftarrow u_i^{(m)} + \alpha^{(m)} \text{CS} (1 - \omega) U_S^{(m)} + \omega U_X^{(m)}$$

の形で与えることとした。

学習則と報酬帰属

更新は 2 成分 \$U_S^{(m)}\$ と \$U_X^{(m)}\$ に分解される。前者は \$u_i^{(m)}\$ の自己調節によってどの程度報酬が改善したかを表し、後者は獲得期から消去期への移行など、外因的要因による身体状態偏差の変化に対応する成分である。自己帰属成分 \$U_S^{(m)}\$ は、試行間の総報酬変化 \$\Delta R_i\$ を \$u_i^{(m)}\$ の増分で規格化し、更新量を飽和させる非線形関数 \$h\$ を通して

$$U_S^{(m)} = h \left(\left| \frac{\Delta R_i}{\Delta u_i^{(m)} + \varepsilon} \right| \right),$$

と定義した。ここで

$$h(\cdot) = \tanh(\cdot), \quad \Delta R_i = R_i - R_{i-1}, \quad \Delta u_i^{(m)} = u_i^{(m)} - u_{i-1}^{(m)},$$

であり、\$\varepsilon\$ はゼロ除算回避のための微小定数である。外因帰属成分 \$U_X^{(m)}\$ は、報酬変化を偏差量 \$X_i\$ の変化で規格化して

$$U_X^{(m)} = h \left(\left| \frac{\Delta R_i}{\Delta X_i + \varepsilon} \right| \right), \quad \Delta X_i = X_i - X_{i-1},$$

と定義した。両者の混合比は \$\omega \in [0,1]\$ により決まり、

$$\omega = \sigma \left(h \left(\left| \frac{\Delta R_i}{\Delta s_i + \varepsilon} \right| \right) - c \right),$$

とする。ここで \$\sigma(\cdot) = 1/(1 + \exp(\cdot))\$ はシグモイド関数、\$\Delta s_i = s_i - s_{i-1}\$、\$c\$ は閾値パラメータである。直観的には、\$\omega\$ は \$R_i\$ の変化が自己調節に由来する可能性と、外因的攪乱に由来する可能性を表す指標である（図 S1）。\$R_i\$ が小さな更新に伴って緩やかに改善する場合には \$\omega\$ は小さく、\$U_S^{(m)}\$ が支配的となる。逆に、身体状態の予期せぬ逸脱により \$R_i\$ が急変する場合には \$\omega\$ が大きくなり、\$U_X^{(m)}\$ に基づく更新が優勢となる。このとき、\$R_i\$ の変化は \$u_i^{(m)}\$ に直接対応しにくいいため、\$u_i^{(m)}\$ の更新に用いる勾配は、偏差量 \$X_i\$ に対する有限差分により近似される。

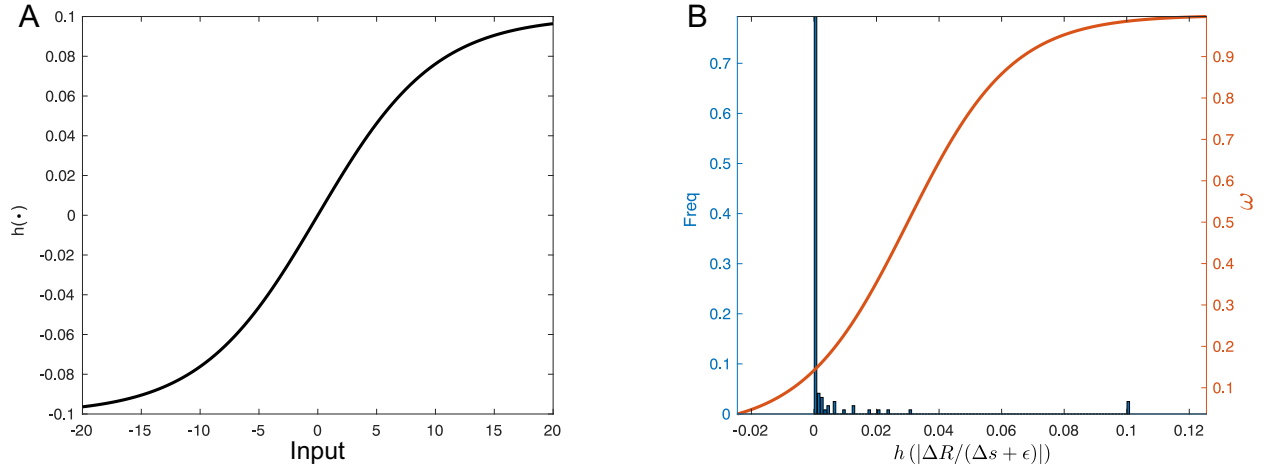


図 S1. 代表データセットに対する $h(x) = C_1 \tanh(C_2 x)$ および $\sigma(x) = 0.5 + C_3 \tanh(C_4 x)$. (A) $U_S^{(m)}$ および $U_X^{(m)}$ を算出する際に用いる閾値化関数 h . (B) 左：閾値化関数 h により閾値化されたサンプル観測値の出現頻度。右：シグモイド関数 σ (橙線)。これは環境遷移（外因的変化）の確率 ω を算出する。座標軸およびパラメータ設定は本文の解析で用いたものと一致させてある。シミュレーションは全 120 試行からなり、1–81 試行が獲得期、82–90 試行が消去期、91–120 試行が再獲得期に対応する。

更新信号の強化学習としての解釈

本モデルは連続時間で実装され、神経回路様のダイナミクスをもつ一方、学習は試行単位で進行し、強化学習の枠組みで解釈できる。中心となる学習変数は更新信号 U_S と U_X であり、総報酬の試行間差分（試行レベルの報酬予測誤差）を、予期的自己調節に帰属すべき成分と外因的攪乱に帰属すべき成分へ分離する。これらの差分は正規化され、興奮性・抑制性成分へ配分されることで、ノイズに起因する符号反転や不安定な更新を抑制する。得られた更新信号に基づき、将来の補償応答を規定する予期制御変数が調整される。したがって、提案モデルの学習は試行ごとの強化学習過程として理解できる。

アルゴリズム的観点から見ると、本学習則は Q 学習や方策勾配法などの標準的名称をもつ強化学習アルゴリズムには直接対応しない。むしろ、総報酬の有限差分を用いて、予期制御変数に関する報酬勾配を構造化された形で近似する、試行レベルの勾配型報酬最大化スキームを実装している。この枠組みは、行動（予期制御信号）が報酬結果のみで評価され、状態遷移モデルを明示的に用いないという点でバンディット型学習に近い。ただし通常のバンディットと異なり、クレジット割り当てがホメオスタシス動態と内部状態変数によって形作られるため、生物学的妥当性をもつ勾配近似として位置付けられる。

A1.2 基準 HRL モデルの実装 (Keramati & Gutkin, 2014)

本節では、Keramati & Gutkin (2014) の原典モデルに対応する基準実装を示す。これは比較のための参照モデルであり、状態ダイナミクスおよび報酬計算は提案モデルと同一である一方、制御の表現と更新方法が異なる。具体的には、予期制御離散行動集合から価値ベース強化学習により選択される。

試行 i で得られる総報酬は、瞬時報酬の時間割引付き積分として

$$R_i = \int_0^{T_i} \gamma(t) r_i(t) dt, \quad \gamma(t) = \exp(-\lambda t),$$

と定義する。ここで T_i は試行 i の持続時間、 $\lambda > 0$ は割引率である。瞬時報酬は

$$r_i(t) = -\frac{dD}{dt}$$

で与え、状態変化の方向がホメオスタシス・ドライブとどの程度整合するかを定量化する。基準モデルは、離散行動 $a \in \{0,1\}$ に対する価値関数 $Q(a)$ を保持し、試行 i の後、学習率を α として、選択された行動 A_i の価値のみを

$$Q(A_i) \leftarrow Q(A_i) + \alpha(R_i - v(A_i))$$

により更新する（非選択行動の価値は更新しない）。次試行の行動選択は softmax 方策

$$P(A_{i+1} = a) = \frac{\exp(Q(a))}{\sum_{a' \in \{0,1\}} \exp(Q(a'))}$$

に従う。以上より、本モデルは二腕バンディット型の強化学習として解釈できる。

A1.3 感度解析

本研究の一次元モデルは、反応的制御項、外乱項、予期的補償応答、および報酬に基づく学習更新から構成される連続時間力学系である。モデルには学習率や抑制強度、報酬の非対称性、外乱帰属の非線形関数など複数のパラメータが含まれるため、主要な定性的結論（獲得・消去・再獲得の相構造、逸脱の抑制、外乱帰属の振る舞い）が、妥当なパラメータ範囲で頑健に再現されるかを確認する必要がある。そこで一次元モデルについて感度解析を行い、どのパラメータがどの指標に強く影響するかを定量化した。シミュレーションは全 120 試行からなり、1-81 試行が獲得期、82-90 試行が消去期、91-120 試行が再獲得期に対応する。

A1.3.1 one-at-a-time (OAT) 法

感度解析には、計算負荷が小さく解釈が明快な one-at-a-time (OAT) 法を用いた。OAT 法では、基準パラメータ集合 θ_0 を固定したまま、ある 1 つのパラメータ p のみを $\pm\delta$ （本研究では $\delta = 0.1, 0.2, 0.5$ の 3 条件）だけ変化させ、出力指標の変化を評価する。具体的には、 p を $p(1+\delta)$ および $p(1-\delta)$ に変更した 2 条件のシミュレーションを行い、指標 M に対する正規化感度係数 $S_{M,p}$ を

$$S_{M,p} = \frac{M(p(1+\delta)) - M(p(1-\delta))}{2\delta \max(|M(p)|, \varepsilon_M)}$$

で定義した。ここで ε_M はゼロ除算を避けるための小定数である。 $S_{M,p}$ の符号は、パラメータを増加させたときに指標が増減する方向を示し、絶対値 $|S_{M,p}|$ は影響の大きさを表す。OAT は局所感度の手法であるが、まず主要パラメータの相対的重要度を把握する目的には十分であり、広く用いられる。

対象としたパラメータは、一次元モデルの挙動に直接関与し、かつ生理学的・計算論的解釈が明確なものとして下記を採用した（モデル詳細は補遺 A1.1, 基準値は表 S1 を参照）。

- 報酬の非対称性 η ：悪化側の報酬重み（ドライブ増加のペナルティ）
- 抑制強度 $w^{A,I}$ ：抑制学習成分の寄与（ u_I の有効強度）
- 学習率 α^A, α^I ：促進・抑制成分の更新速度
- 自己帰属更新の非線形 $h(x) = \tanh(b_h x)$ の傾きパラメータ b_h ： $\Delta R / \Delta s$ に基づく更新信号の飽和特性
- 外乱帰属関数 $\sigma(x) = \frac{1}{2} + \tanh\left(b_\omega(x + c_\omega)\right)$ の形状パラメータ：自己由来更新と外因由来更新の混合比を決める関数の感度（傾き b_ω , 閾値 c_ω ）

なお、 α^I と $w^{A,I}$ はともに抑制項 $w^{A,I}u_I$ の有効強度をスケールするため、局所領域では類似の影響を示しうる（識別性の観点からは、両者が同時に変化する場合の効果は積により部分的に縮約される）。

今回の一次元モデルの感度解析で対象から外した係数 a はドライブ関数 $D(x) = a(x - x^*)^2$ のスケールを定め、反応的ホメオスタシス項 $-dD/dx$ の復元力（戻りの速さ）を直接規定するだけでなく、報酬定義 $r(t)$ （ドライブ変化に基づく評価）の振幅にも影響する。そのため a は、学習更新の大きさや閾値化・飽和関数が動作するダイナミックレンジとも相互作用しうる、いわば“スケール支配パラメータ”である。一次元モデルでは、実験データの

体温変動スケール（図 2）を再現することを優先し、 a を基準設定として固定した上で、上記の学習・抑制・帰属に関わるパラメータの効果を分離して評価した。これに対し多次元拡張では、共有制御の下で複数の恒常性変数が競合するため、スケール設定（復元力と報酬スケール）がトレードオフ構造や学習の成功条件に与える影響が相対的に大きくなる。以上の理由から、多次元モデルでは a を系統的に検討した。

感度解析で用いる指標 M は、学習過程の相（獲得・消去・再獲得）を反映し、かつモデルの目的である「逸脱の抑制」と「帰属（自己/外因）の推定」を捉えるものとして以下を採用した。

1. 相別平均報酬

試行 i の報酬 R_i を用い、獲得期（acquisition）、消去期（extinction）、再獲得期（retrieval）それぞれの平均 $R_{\text{acq}} = \langle R_i \rangle_{i \in \text{acq}}$, $R_{\text{ext}} = \langle R_i \rangle_{i \in \text{ext}}$, $R_{\text{ret}} = \langle R_i \rangle_{i \in \text{ret}}$ を計算した。各相の区切りは US 消去区間 ($t_{\text{US,extinc}}$) の設定に基づき、シミュレーション内の試行インデックスを割合で分割して定義した。

2. 累積逸脱量

各試行での偏差の時間積分 $X_i = \int (x(t) - x^*) dt$ を用い、全試行にわたる累積逸脱

$$X_{\text{tot}} = \sum_i |X_i|$$

を計算した（値が小さいほど恒常性が良く維持される）。

3. 外乱帰属の平均（自己／外因の混合比）

試行 i における外乱帰属係数 $\omega_i \in [0,1]$ を、自己由来更新と外因由来更新の混合比として定義し、その試行平均

$$\bar{\omega} = \langle \omega_i \rangle_i$$

を評価した。

結果を表 S1 に示す。OAT の結果、罰の非対称性 η は、相別平均報酬および累積逸脱量に対して大きな影響を持ち、一次元モデルの学習ダイナミクスと恒常性維持性能を強く規定する主要因であることが示された。一方、外乱帰属関数 $\sigma(\cdot)$ の形状パラメータは、主として $\bar{\omega}$ を変化させる（すなわち帰属判断そのものへ影響する）傾向が見られ、性能指標への影響は局所摂動では相対的に小さく見える場合があった。これは、 $\sigma(\cdot)$ の閾値型の非線形性により、入力分布が閾値近傍にある場合にのみ帰属判断が急峻に切り替わるためである。そこで次節では、 ω の閾値パラメータを広い範囲でスイープし、帰属変化が学習挙動に与える影響を直接可視化する。

表 S1. 一次元モデルの感度解析 (OAT) の結果. 上段から $\delta = 0.2, 0.1, 0.5$. パラメータはそれぞれ、 η : 報酬の非対称性、 $w^{A,I}$: 抑制強度、 α^A, α^I : 学習率 (促進・抑制成分)、 b_h : 自己帰属更新の非線形 $h(\cdot)$ の傾き、 b_ω, c_ω : 自己由来更新と外因由来更新の混合比を決める関数の感度 $\sigma(\cdot)$ の形状パラメータ (それぞれ傾きと閾値)。

$\delta = 0.2$

Parameter	p_0	p_-	p_+	$S_{R_{\text{acq}}}$	$S_{R_{\text{ext}}}$	$S_{R_{\text{ret}}}$	$S_{X_{\text{tot}}}$	$S_{\bar{\omega}}$	Score
η	1.5	1.2	1.8	-2.6629	-2.8439	-4.6310	0.9647	0.1722	3.6275
b_h	0.1	0.08	0.12	0.10621	-0.0202	-0.7855	0.6031	0.0103	0.7893
α^A	1.5	1.2	1.8	0.21913	0.0202	-0.5210	0.3347	-0.0781	0.5538
α^I	4	3.2	4.8	0	0.1027	-0.5392	0.4621	0.0036	0.4621
$w^{A,I}$	1	0.8	1.2	0	0.1027	-0.5392	0.4621	0.0036	0.4621
b_ω	30	24	36	-0.0078	0.0208	0.4869	-0.3761	-1.2603	0.3838
c_ω	-0.03	-0.024	-0.036	-0.0651	0.1140	0.1446	-0.1187	-1.3211	0.1837

$\delta = 0.1$

Parameter	p_0	p_-	p_+	$S_{R_{\text{acq}}}$	$S_{R_{\text{ext}}}$	$S_{R_{\text{ret}}}$	$S_{X_{\text{tot}}}$	$S_{\bar{\omega}}$	Score
η	1.5	1.35	1.65	-2.6944	-3.0395	-4.7277	1.3434	0.3369	4.0378
b_h	0.1	0.09	0.11	0.1006	-0.0501	-2.9313	1.8044	0.2213	1.9050
α^A	1.5	1.35	1.65	0.2120	0.0286	-0.5032	0.3111	-0.3023	0.5231
α^I	1	0.9	1.1	0	0.0791	-0.5865	0.5176	-0.0151	0.5176
$w^{A,I}$	4	3.6	4.4	0	0.0791	-0.5865	0.5176	-0.0155	0.5176
c_ω	-0.03	-0.027	-0.033	-0.0652	0.1054	0.4624	-0.2541	-1.4501	0.3193
b_ω	30	27	33	-0.0083	0.0257	0.2098	-0.1958	-1.3742	0.2041

$\delta = 0.5$

Parameter	p_0	p_-	p_+	$S_{R_{\text{acq}}}$	$S_{R_{\text{ext}}}$	$S_{R_{\text{ret}}}$	$S_{X_{\text{tot}}}$	$S_{\bar{\omega}}$	Score
η	1.5	1	2.25	-2.4845	-3.2830	-44.2394	-0.4641	1.2271	2.9485
b_h	0.1	0.05	0.15	0.1491	0.1034	-0.7211	0.5254	0.1258	0.6745
α^A	1.5	0.75	2.25	0.2849	0.0179	-0.4233	0.1898	-0.1871	0.4747
α^I	1	0.5	1.5	0	0.1869	-0.7400	0.4509	0.0145	0.4509
$w^{A,I}$	4	2	6	0	0.1869	-0.7400	0.4509	0.0151	0.4509
c_ω	-0.03	-0.015	-0.045	-0.0645	0.1095	0.4025	-0.3147	-1.2888	0.3792
b_ω	30	15	45	-0.0033	0.0156	0.1881	-0.1778	-1.1977	0.1811

A1.3.2 報酬の自己帰属・外因帰属についての議論のための補足 (ω のスイープ)

前節の OAT は主要パラメータの相対的重要度を整理するのに有効である一方、 ω に関しては報酬・状態の変動の自己帰属・外因帰属の解釈を議論するため、 ω の変化が学習ダイナミクスに与える影響を別途可視化した。特に ω の閾値パラメータ c_ω は、入力分布が閾値近傍にある場合に帰属判断を急峻に切り替えるため、乗算型の微小摂動（例： $\pm 10\text{--}20\%$ ）では影響が小さく見えることがある。そこで 閾値パラメータ c_ω を加算的にスイープ ($c_\omega = c_0 + \Delta c$) し、 $\bar{\omega}$ および再獲得期の平均報酬 R_{ret} の変化を評価した（図 S2）。外乱帰属係数 $\omega_i \in [0,1]$ は、試行 i において報酬変化が自己由来更新と外因由来更新のどちらに帰属されるかを表す混合比である。とくに閾値パラメータ c_ω は、帰属判断を切り替える境界を規定するため、乗算的な微小摂動よりも、閾値を跨ぐ加算的スイープによって影響が明確になる。そこで 閾値パラメータ c_ω を $c_\omega = c_0 + \Delta c$ として加算的にスイープし、外乱帰属の平均 $\bar{\omega}$ と再獲得期の平均報酬 R_{ret} の変化を評価した（図 S2）。

その結果、 $\bar{\omega}$ が増加（外因帰属が優勢化）するにつれて R_{ret} が悪化し、学習された予期的補償応答の形成あるいは再獲得が阻害されることが確認された。すなわち、外乱を外因性として強く推定する条件では、自己調節に基づく更新が抑制され、内部状態と有効な代償反応の連合形成が進みにくいことが見てとれた。

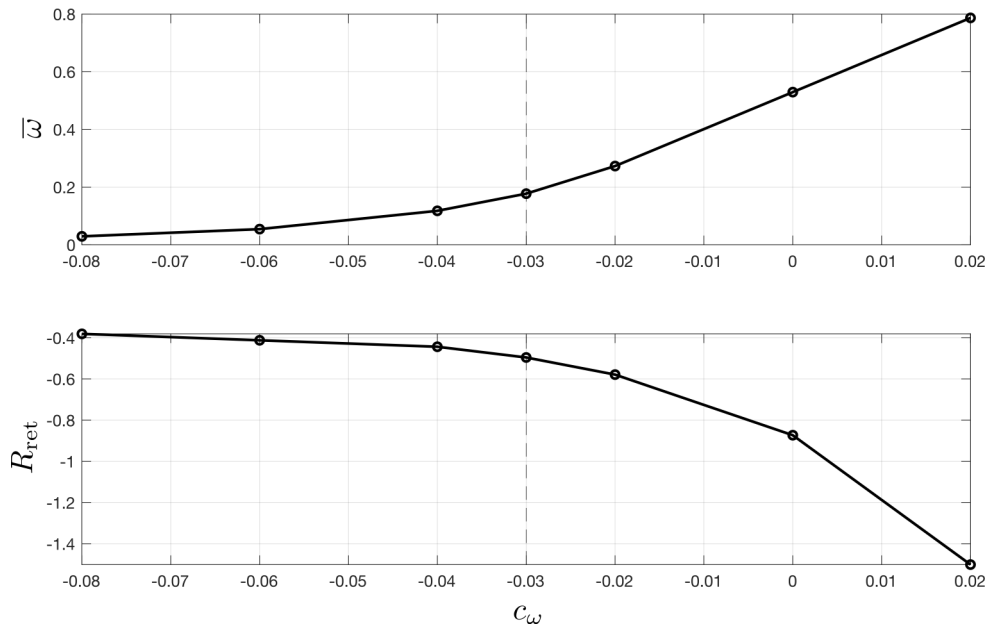


図 S2. 外乱帰属の閾値パラメータ c_ω のスイープによる帰属と再獲得性能の変化. 外乱帰属確率の試行平均 $\bar{\omega} = \langle \omega_i \rangle_i$ (上段) および再獲得期の平均報酬 R_{ret} (下段)。破線は基準設定 $c_\omega = c_0$ を示す。

A2. 多次元拡張モデル

一次元モデルでは、体温という単一の内部変数の調節を扱った。しかし実際の生体は、複数の生理変数について同時にホメオスタシスを維持する必要がある。本節では、脳が末梢神経活動やホルモン分泌などの限られた指令信号を通じて、複数の内部変数（例：体温、血圧、血糖）を制御する状況を考える。この場合、各指令信号は複数の変数へ同時に影響を及ぼすため、ある変数を最適値へ近づける調整が、他の変数を基準値から逸脱させることがあり得る。例えば、発汗による体温低下は体温を適正域へ戻す一方で体内水分を失わせる。薬理介入でも同様に、ある症状を緩和しても別の生理次元を攪乱する副作用が生じ得る。ホメオスタシス強化学習（HRL）の枠組みでは、各内部変数は固定された基準値をもち、ある行動が一つの変数を最適化しても、他の変数を必ずしも基準値へ近づけるとは限らない。したがって調節行動は、利益と不利益を同時にもたらす不可避なトレードオフを伴う。

試行 i において、各変数 $x_n(t)$ は局所信号 $s_{n,i}$ により調節され、これはメタ信号 $z_{k,i}$ の加重和として

$$s_{n,i} = \sum_{k=1}^K w_{kn} z_{k,i},$$

で与えられる。ここで w_{kn} は k 番目メタ信号から n 番目変数への固定の生理学的影響重みである。 n 番目変数のダイナミクスは

$$\frac{\partial x_{n,i}}{\partial t} = \frac{\partial D_n}{\partial x_{n,i}} + F_n(t - t_{US,i}) + s_{n,i} G_n(t - t_{CS,i}),$$

と定義した。ここで

$$D_n(x_n) = a_n(x_n - x_n^*)^2$$

はドライブ関数、 $t_{US,i}$ と $t_{CS,i}$ はそれぞれ US と CS の提示時刻、 $F_n(t)$ と $G_n(t)$ は n 番目変数に対する US および CS 誘発入力関数である。 $G_n(t)$ の振幅は局所信号 $s_{n,i}$ により調節され、 $s_{n,i}$ 自体がメタ信号によって制御される。内部状態は基準値からの偏差として表現し、正負は基準値より上・下への逸脱を示す。瞬時報酬 $r_n(t)$ は変数ごとに独立に定義し、一次元モデルと同様に総報酬 $R_{n,i}$ と偏差の累積量 $X_{n,i}$ を計算する。

$$r_n(t) = \begin{cases} -\frac{\partial D_n}{\partial t} & \text{if } \frac{\partial D_n}{\partial t} \leq 0 \\ -\eta \frac{\partial D_n}{\partial t} & \text{if } \frac{\partial D_n}{\partial t} > 0 \end{cases}$$

$$R_{n,i} = \int_{t_{CS,i}}^{T_i} r_{n,i}(t) dt, \quad X_{n,i} = \int_{t_{CS,i}}^{T_i} (x_n(t) - x_n^*(t)) dt.$$

メタ信号 z_k は制御変数 u_k により最適化され、全変数の総報酬 $\sum_{n=1}^N R_n$ を最大化する。メタ信号 z_k は、興奮性集団 A_k と抑制性集団 I_k により調節され、 $m \in \{A, I\}$ に対する制御変数 $u_k^{(m)}$ が更新される。具体的には、

$$z_{k,i} = b(A_{k,i} - \chi_i w^{A,I} I_{k,i}), \quad A_{k,i} = u_{k,i}^A \text{CS}, \quad I_{k,i} = u_{k,i}^I \text{CS},$$

とする。ゲーティング変数 χ_i は直前試行に基づいて導入し (A1.1 参照)、 $w^{A,I}$ は抑制効果の強さ、 b は神経活動が状態ダイナミクスへ及ぼすスケーリング係数を表す。更新は、自己調節に基づく $U_{Z_n}^{(m)}$ と、外乱・偏差に基づく $U_{X_n}^{(m)}$ の 2 成分からなり、

$$U_{Z_n}^{(m)} = h \left(\left| \frac{\Delta R_n}{\Delta u_k^{(m)} + \varepsilon} \frac{\partial z_k}{\partial u_k^{(m)}} \right| \right), \quad U_{X_n}^{(m)} = h \left(\left| \frac{\Delta R_n}{\Delta X_n + \varepsilon} \frac{\partial X_n}{\partial u_k^{(m)}} \right| \right),$$

と定義した。ここで $\Delta R_n = R_{n,i} - R_{n,i-1}$ 、 $\Delta X_n = X_{n,i} - X_{n,i-1}$ 、 ε はゼロ除算回避の微小定数である。 $\partial X_n / \partial u_k^{(m)}$ は、状態ダイナミクスを介して制御変数の変化が累積偏差へ及ぼす有効感度を表す。両成分の相対寄与は変数ごとの重み ω_n により

$$\omega_n = \sigma \left(h \left| \frac{\Delta R_n}{\Delta z_k + \varepsilon} \right| - c \right)$$

で与えた。ここで c はシグモイドの位置を調整するパラメータである。全パラメータは表 1 にまとめた。

優先度付けのパラメータ化とパラメータスweepについての補足

優先度重み $a = (a_1, a_2, a_3)$ は、単一の角度変数 $\theta \in [0, 2\pi)$ により

$$a_1 = \beta(\cos \theta + 1), \quad a_2 = \beta \left(\cos \left(\theta + \frac{2\pi}{3} \right) + 1 \right), \quad a_3 = \beta \left(\cos \left(\theta + \frac{4\pi}{3} \right) + 1 \right),$$

とパラメータ化した。この定義により、優先度重みは常に非負となり、3 変数間で優先度が対称的に回転する。概ね均衡した優先度は a_n が互いに近い値を取る設定に対応し、 θ の極端な値では単一変数が強く優先される不均衡な優先度付けが実現される。

先行モデルのパラメータ m, n と提案モデルの a, u の関係

既存モデルは本質的に多次元であり、自由パラメータ m, n がドライブ関数の基盤を成す。現行のドライブ関数の定義では、各次元の逸脱 ($x_i^* - x_{i,t}$) を独立に扱い、それらを総和した上で m 乗することで全体の動機づけを定める。すなわち、各次元の逸脱の線形和として全体の動機づけを解釈し、次元間相互作用は考慮しない。

これに対し、提案モデルでは状態ごとにドライブを算出し、状態間の優先度をパラメータ a により規定する。さらに、制御信号の更新に関与する方策側のパラメータを u として明示し、優先度 a （何をどれだけ重視するか）と、学習により変化する制御戦略 u （どのように調節するか）を分離して扱う構成とした。これにより、優先度配分の効果と学習ダイナミクスを明確に切り分けつつ、変数間のトレードオフと収束特性を解析できるモデルとした。

以下にパラメータの具体的な値を示す。

表 S2. パラメータ値

Parameter	Value
T_{trial}	50
α^A	1.5
α^I	4
α	0.3
b	2
$w^{A,I}$	1
β	2
c	-0.03
C_1	0.1
C_2	0.1
C_3	0.5
C_4	30
λ	0.99

文献

- Annesley, S. J., Missailidis, D., Heng, B., Josev, E. K., & Armstrong, C. W. (2024). Unravelling shared mechanisms: insights from recent ME/CFS research to illuminate long COVID pathologies. *Trends in Molecular Medicine*, 30(5), 443–458.
- Arroll, M. A. (2013). Allostatic overload in myalgic encephalomyelitis/chronic fatigue syndrome (ME/CFS). *Medical Hypotheses*, 81(3), 506–508.
- Ashby, W. (1960). *Design for a brain: The origin of adaptive behaviour*. Chapman & Hall.
- Barrett, L. F., & Simmons, W. K. (2015). Interoceptive predictions in the brain. *Nature Reviews. Neuroscience*, 16(7), 419–429.
- Berridge, K. C., Robinson, T. E., & Aldridge, J. W. (2009). Dissecting components of reward: “liking”, “wanting”, and learning. *Current Opinion in Pharmacology*, 9(1), 65–73.
- Bouton, M. E. (2002). Context, ambiguity, and unlearning: sources of relapse after behavioral extinction. *Biological Psychiatry*, 52(10), 976–986.
- Bouton, M. E., Westbrook, R. F., Corcoran, K. A., & Maren, S. (2006). Contextual and temporal modulation of extinction: behavioral and biological mechanisms. *Biological Psychiatry*, 60(4), 352–360.
- Cannon, W. B. (1932). The wisdom of the body. *The American Journal of the Medical Sciences*, 184(6), 864.
- Clark, J. J., & Bernstein, I. L. (2006). Sensitization of salt appetite is associated with increased “wanting” but not “liking” of a salt reward in the sodium-deplete rat. *Behavioral Neuroscience*, 120(1), 206–210.
- Craig, A. D. (2002). How do you feel? Interoception: the sense of the physiological condition of the body. *Nature Reviews. Neuroscience*, 3(8), 655–666.

- Craig, A. D. (2003). A new view of pain as a homeostatic emotion. *Trends in Neurosciences*, 26(6), 303–307.
- Escorihuela, R. M., Capdevila, L., Castro, J. R., Zaragoza, M. C., Maurel, S., Alegre, J., & Castro-Marrero, J. (2020). Reduced heart rate variability predicts fatigue severity in individuals with chronic fatigue syndrome/myalgic encephalomyelitis. *Journal of Translational Medicine*, 18(1), 4.
- Farahbakhsh, Z. Z., & Siciliano, C. A. (2023). Pavlovian-conditioned opioid tolerance. *Science Advances*, 9(6), eadg6086.
- Freeman, R., & Komaroff, A. L. (1997). Does the chronic fatigue syndrome involve the autonomic nervous system? *The American Journal of Medicine*, 102(4), 357–364.
- Friston, K. (2013). Life as we know it. *Journal of the Royal Society, Interface*, 10(86), 20130475.
- Friston, K. J., & Stephan, K. E. (2007). Free-energy and the brain. *Synthese*, 159(3), 417–458.
- Friston, K., Levin, M., Sengupta, B., & Pezzulo, G. (2015). Knowing one's place: a free-energy approach to pattern regulation. *Journal of the Royal Society, Interface / the Royal Society*, 12(105). <https://doi.org/10.1098/rsif.2014.1383>
- Ganzel, B. L., Morris, P. A., & Wethington, E. (2010). Allostasis and the human brain: Integrating models of stress from the social and life sciences. *Psychological Review*, 117(1), 134–174.
- Garner, R., & Baraniuk, J. N. (2019). Orthostatic intolerance in chronic fatigue syndrome. *Journal of Translational Medicine*, 17(1), 185.
- Goertzel, B. N., Pennachin, C., de Souza Coelho, L., Maloney, E. M., Jones, J. F., & Gurbaxani, B. (2006). Allostatic load is associated with symptoms in chronic fatigue syndrome patients. *Pharmacogenomics*, 7(3), 485–494.

- Guyton, A. C., Jones, C. E., & Coleman, T. G. (1973). Circulatory Physiology: Cardiac Output and its Regulation. *Annals of Internal Medicine*, 79(5), 768.
- Herry, C., Ciocchi, S., Senn, V., Demmou, L., Müller, C., & Lüthi, A. (2008). Switching on and off fear by distinct neuronal circuits. *Nature*, 454(7204), 600–606.
- Hou, Y., Zou, G., Wang, X., Guo, H., Ma, X., Cheng, X., Xie, Z., Zuo, X., Xia, J., Mao, H., Yuan, M., Chen, Q., Cao, P., Yang, Y., Zhang, L., & Xiong, W. (2023). Coordinated activity of a central pathway drives associative opioid analgesic tolerance. *Science Advances*, 9(6), eabo5627.
- Hull, C. L. (1943). *Principles Of Behavior: an introduction to behavior theory* (R. M. Elliott, Ed.). Appleton-Century.
- Iwashyna, T. J. (2012). Trajectories of recovery and dysfunction after acute illness, with implications for clinical trial design [Review of *Trajectories of recovery and dysfunction after acute illness, with implications for clinical trial design*]. *American Journal of Respiratory and Critical Care Medicine*, 186(4), 302–304. American Thoracic Society.
- Keramati, M., Durand, A., Girardeau, P., Gutkin, B., & Ahmed, S. H. (2017). Cocaine addiction as a homeostatic reinforcement learning disorder. *Psychological Review*, 124(2), 130–153.
- Keramati, M., & Gutkin, B. (2014). Homeostatic reinforcement learning for integrating reward collection and physiological stability. *ELife*, 3. <https://doi.org/10.7554/eLife.04811>
- Le, A., Poulos, C., & Cappell, H. (1979). Conditioned tolerance to the hypothermic effect of ethyl alcohol. In *Science* (Vol. 206, Issue 4422, pp. 1109–1110). <https://doi.org/10.1126/science.493999>

- Li, Y., Nakae, K., Ishii, S., & Naoki, H. (2016). Uncertainty-Dependent Extinction of Fear Memory in an Amygdala-mPFC Neural Circuit Model. *PLoS Computational Biology*, *12*(9), e1005099.
- Likhtik, E., Popa, D., Apergis-Schoute, J., Fidacaro, G. A., & Paré, D. (2008). Amygdala intercalated neurons are required for expression of fear extinction. *Nature*, *454*(7204), 642–645.
- Livneh, Y., Ramesh, R. N., Burgess, C. R., Levandowski, K. M., Madara, J. C., Fenselau, H., Goldey, G. J., Diaz, V. E., Jikomes, N., Resch, J. M., Lowell, B. B., & Andermann, M. L. (2017). Homeostatic circuits selectively gate food cue responses in insular cortex. *Nature*, *546*(7660), 611–616.
- Macpherson, T., Morita, M., & Hikida, T. (2014). Striatal direct and indirect pathways control decision-making behavior. *Frontiers in Psychology*, *5*, 1301.
- Mansfield, J. G., Benedict, R. S., & Woods, S. C. (1983). Response specificity of behaviorally augmented tolerance to ethanol supports a learning interpretation. *Psychopharmacology*, *79*(2–3), 94–98.
- Mansfield, J. G., & Cunningham, C. L. (1980). Conditioning and extinction of tolerance to the hypothermic effect of ethanol in rats. *Journal of Comparative and Physiological Psychology*, *94*(5), 962–969.
- McEwen, B. S. (2017). Stress: Homeostasis, Rheostasis, Reactive Scope, Allostasis and Allostatic Load☆. In *Reference Module in Neuroscience and Biobehavioral Psychology*. Elsevier.
- McEwen, B. S., & Stellar, E. (1993). Stress and the individual. Mechanisms leading to disease. *Archives of Internal Medicine*, *153*(18), 2093–2101.
- McEwen, B. S., & Wingfield, J. C. (2003). The concept of allostasis in biology and biomedicine. *Hormones and Behavior*, *43*(1), 2–15.

- Montague, P. R., Dayan, P., & Sejnowski, T. J. (1996). A framework for mesencephalic dopamine systems based on predictive Hebbian learning. *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience*, 16(5), 1936–1947.
- Myers, K. M., & Davis, M. (2002). Behavioral and neural analysis of extinction. *Neuron*, 36(4), 567–584.
- Napier, R. M., Macrae, M., & Kehoe, E. J. (1992). Rapid reacquisition in conditioning of the rabbit's nictitating membrane response. *Journal of Experimental Psychology. Animal Behavior Processes*, 18(2), 182–192.
- Nelson, M. J., Bahl, J. S., Buckley, J. D., Thomson, R. L., & Davison, K. (2019). Evidence of altered cardiac autonomic regulation in myalgic encephalomyelitis/chronic fatigue syndrome: A systematic review and meta-analysis. *Medicine*, 98(43), e17600.
- Pezzulo, G., Parr, T., & Friston, K. (2024). Active inference as a theory of sentient behavior. *Biological Psychology*, 186, 108741.
- Poulos, C. X., & Cappell, H. (1991). Homeostatic theory of drug tolerance: A general model of physiological adaptation. *Psychological Review*, 98(3), 390–408.
- Quirk, G. J., & Mueller, D. (2008). Neural mechanisms of extinction learning and retrieval. *Neuropsychopharmacology: Official Publication of the American College of Neuropsychopharmacology*, 33(1), 56–72.
- Raizen, D. M., Mullington, J., Anaclet, C., Clarke, G., Critchley, H., Dantzer, R., Davis, R., Drew, K. L., Fessel, J., Fuller, P. M., Gibson, E. M., Harrington, M., Ian Lipkin, W., Klerman, E. B., Klimas, N., Komaroff, A. L., Koroshetz, W., Krupp, L., Kuppaswamy, A., ... Craig Heller, H. (2023). Beyond the symptom: the biology of fatigue. *Sleep*, 46(9). <https://doi.org/10.1093/sleep/zsad069>

- Ramsay, D. S., & Woods, S. C. (2014). Clarifying the roles of homeostasis and allostasis in physiological regulation. *Psychological Review*, 121(2), 225–247.
- Rescorla, R. A., & Heth, C. D. (1975). Reinstatement of fear to an extinguished conditioned stimulus. *Journal of Experimental Psychology. Animal Behavior Processes*, 1(1), 88–96.
- Rescorla, Robert A., & Wagner, A. R. (1972). *Theory Pavlovian conditioning: Variations effectiveness reinforcement non reinforcement*.
- Schultz, W., Dayan, P., & Montague, P. R. (1997). A neural substrate of prediction and reward. *Science (New York, N.Y.)*, 275(5306), 1593–1599.
- Seth, A. K. (2013). Interoceptive inference, emotion, and the embodied self. *Trends in Cognitive Sciences*, 17(11), 565–573.
- Sharp, H., Themelis, K., Amato, M., Barritt, A., Davies, K., Harrison, N., Critchley, H., Garfinkel, S., & Eccles, J. (2021). The role of interoception in the mechanism of pain and fatigue in fibromyalgia and myalgic encephalomyelitis/chronic fatigue syndrome (ME/CFS). *European Psychiatry: The Journal of the Association of European Psychiatrists*, 64(S1), S139–S139.
- Siegel, S. (2001). Pavlovian Conditioning and Drug Overdose: When Tolerance Fails. In *Addiction Research & Theory* (Vol. 9, Issue 5, pp. 503–513).
<https://doi.org/10.3109/16066350109141767>
- Siegel, S. (2005). Drug tolerance, drug addiction, and drug anticipation. *Current Directions in Psychological Science*, 14(6), 296–300.
- Sokol-Hessner, P., & Rutledge, R. B. (2019). The psychological and neural basis of loss aversion. *Current Directions in Psychological Science*, 28(1), 20–27.
- Stephan, K. E., Manjaly, Z. M., Mathys, C. D., Weber, L. A. E., Paliwal, S., Gard, T., Tittgemeyer, M., Fleming, S. M., Haker, H., Seth, A. K., & Petzschner, F. H. (2016).

- Allostatic self-efficacy: A metacognitive theory of dyshomeostasis-induced fatigue and depression. *Frontiers in Human Neuroscience*, 10, 550.
- Sterling, P. (2004). Principles of allostasis: Optimal design, predictive regulation, pathophysiology, and rational therapeutics. In J. Schulkin (Ed.), *Allostasis, Homeostasis, and the Costs of Physiological Adaptation* (pp. 17–64). Cambridge University Press.
- Sterling, P., & Eyer, J. (1988). Allostasis: A New Paradigm to Explain Arousal Pathology. In S. F. A. Reason (Ed.), *Handbook of Life Stress, Cognition and Health* (pp. 629–649). John Wiley & Sons.
- Sutton, R. S., & Barto, A. G. (2018). *Reinforcement Learning* (2nd ed.). Bradford Books.
- Tschantz, A., Barca, L., Maisto, D., Buckley, C. L., Seth, A. K., & Pezzulo, G. (2022). Simulating homeostatic, allostatic and goal-directed forms of interoceptive control using active inference. *Biological Psychology*, 169, 108266.
- Uchida, Y., Hikida, T., & Yamashita, Y. (2022). Computational Mechanisms of Osmoregulation: A Reinforcement Learning Model for Sodium Appetite. *Frontiers in Neuroscience*, 16, 857009.
- Voiriot, G., Oualha, M., Pierre, A., Salmon-Gandonnière, C., Gaudet, A., Jouan, Y., Kallel, H., Radermacher, P., Vodovar, D., Sarton, B., Stiel, L., Bréchet, N., Préau, S., Joffre, J., & la CRT de la SRLF. (2022). Chronic critical illness and post-intensive care syndrome: from pathophysiology to clinical challenges. *Annals of Intensive Care*, 12(1), 58.

謝辞

本学位論文は、筆者が京都大学 大学院生命科学研究科 高次生命科学専攻 理論生物学分野に進学後5年間の研究のうち、自身の主研究として取り組んだ内容についてまとめたものである。研究の遂行にあたり、本当に多くの方に温かいご支援をいただいた。

まず、研究指導教員である名古屋大学 医学系研究科 本田直樹教授（元京都大学 大学院生命科学研究科 高次生命科学専攻 理論生物学分野 准教授、広島大学 大学院統合生命科学研究科 特任教授兼任）には、非常に多岐に渡るご指導とご支援をいただいた。筆者が数理的な研究の背景を持たないにも関わらず博士後期課程から受け入れていただき、基本的なことから根気強くご指導いただいた上、何かアイディアを思いついた際には突飛な話題であっても度々耳を傾け、率直な意見や未熟な考えも含めて面白がって議論にお付き合いくださった。自由闊達に議論できる環境をご提供くださり、のびのびと快適な研究生活を送ることができた。期待を込めて筆者の今後の可能性を信じてくださっているという激励は、自信など簡単に霧散してしまうような状況において、それでも前向きに取り組む励みとなった。また、今後も研究を続けていくにあたって軸となる確固たる取捨選択の指針、そして現実的な戦略など、研究室運営における姿勢についても、非常に多くのことを学ばせていただいた。そして、小さな子どもを持ちながら研究を止めずに修了まで辿り着くことができたのは、ご自身も非常にご多忙の中、生活面についても大変に温かいご理解と柔軟なご配慮をいただいたためである。ここに、心より感謝申し上げる。

京都大学 大学院生命科学研究科 生体制御学分野の松田道行教授、京都大学 大学院生命科学研究科 細胞周期学分野の青木一洋教授には、本田直樹教授の栄転に伴い、在籍の承諾とお力添えをいただき、さらに青木一洋教授には本論文の主査もご快諾いただいた。厚く御礼申し上げます。

京都大学 大学院生命科学研究科 脳機能発達再生制御学分野の今吉格教授には副指導教員、そして副査として本学位論文の審査をお引き受けいただき、京都大学 大学院生命科学研究科 附属生命情報解析教育センターの東樹宏和教授には副査として、京都大学 大学院生命科学研究科 附属生命情報解析教育センターの坂本雅行准教授には専門委員として、審査をご担当いただいた。また、東京大学 大学院薬学系研究科 谷口雄一教授には副指導教員として有意義なご助言をいただいた。大変ご多忙の中ご指導くださった先生方に、感謝の意を表する。

指導教員である本田教授の一つ目の栄転先、広島大学 大学院統合生命科学研究科では、研究室の枠を超えた合同セミナーや勉強会が開かれ、数理生命科学の広いテーマについて触れる機会をいただいた。数理の背景を持たない筆者にとって、その姿勢や方向性、視点を学べる大変貴重な機会であった。藤本仰一教授、栗津暁紀准教授、藤井雅史助教、現筑波大学 生存ダイナミクス研究センター 齋藤 稔准教授には、親身なご指導と温かい激励をいただいた。不慣れな分野で戸惑いもある中、諦めずにやってこられたのは、先生方のご支援のおかげである。各研究室の皆様も含めて、深く感謝申し上げます。

北海道大学 大学院医学研究院 加藤隆弘教授、大阪大学 蛋白質研究所 疋田貴俊教授には、本研究に重要な医学的知見を提供いただいた。ここに深く感謝申し上げます。

九州工業大学 大学院工学研究院 井上雅世准教授には、MIMS 共同研究集会「現象数理学のダイバーシティ」に毎年呼んでいただき、文理の区別を超えた多様な研究分野で活躍される女性研究者と研究やキャリアについて議論する場を提供いただいた。女性率が非常に低いこ

の分野に身を置く中で、方々で活躍、奮闘されている皆様の存在は、実質的な心の支えとなった。心から感謝を申し上げる。

京都大学 生命科学研究科 細胞周期学分野秘書の上杉知子氏、広島大学 統合生命科学研究科事務室の筒本清佳氏、元秘書の平野和枝氏、名古屋大学 医学系研究科 データ駆動生物学分野 秘書の西尾真由美氏には、研究生活で多大なるサポートをいただいた。厚く御礼申し上げます。

京都大学、広島大学、名古屋大学と時間を共にしてきたデータ駆動生物学研究室のスタッフ、学生の皆様とは度々議論させていただき、助言をいただいた。研究自体だけでなく、研究生活における色々についても励ましていただいた。皆様が努力を積み重ねられる姿は、いつも筆者の刺激となった。神経意思決定チームの矢田祐一郎准教授、堤真人特任助教（現 筑波大学 生存ダイナミクス研究センター助教）、大学院生の東野伊織氏、犬塚健剛氏には、毎回有意義な助言をいただき、また多くの刺激を与えていただいた。金子貴輝氏には、度々興味深く挑戦的な話題を提供いただき、楽しく議論する機会をいただいた。ここに書ききれなかった皆様を含め、心から感謝の意を表する。

また、筆者は博士課程在学中に、国立研究開発法人科学技術振興機構 次世代研究者挑戦的研究プログラム支援事業および独立行政法人日本学術振興会に特別研究員として採用いただき、給与と研究費を支援いただいた。厚く御礼申し上げます。

博士課程中の二度の引越しを含め、筆者とその家族の生活を常々支えてくれた義父・博昭、義母・恵美子、母・藤原てる子、そして山あり谷ありの5年間、常に側で励まし伴走してくれた夫・祐一郎、息子・惟智、無理の多い妊娠期間にも関わらず順調な経過のお腹の子に、心より感謝する。

本学位論文は以下の学術論文の内容に基づいて書かれたものである。

Mana Fujiwara and Honda Naoki (2026). “Gradual proactive regulation of body state by reinforcement learning of homeostasis,” *Neuroscience Research*, 223, 105021, DOI: 10.1016/j.neures.2026.105021