意思決定における報酬と好奇心との葛藤の行動データ駆動的解読

(Behavioral data-driven decoding of the mental conflict between reward and curiosity in decision making)

広島大学統合生命科学研究科
数理生命科学プログラム
古仲　裕貴
学生番号：D211498

データ駆動生物学研究室
指導教官：本田直樹　教授

1.要旨

我々は日常において、心の揺れや葛藤に直面しながら意思決定を迫られており、必ずしも合理的に意思決定するとは限らない。些細なものでは、好奇心に従って未だ試したことのないレストランを試すべきか、それとも美味しいことを知っている馴染みのレストランに行くべきかと悩んだりする。また、人生に大きな影響を及ぼすものとして、報酬を求めて高収入で安定した企業に就職すべきか、それとも不安定ながらも夢や好奇心を追い求めるべきかという葛藤もある。そして、これらの葛藤を抱きながら下した決断が常に合理的だとは限らず、時には好奇心に従って報酬に繋がらない行動をとったりもする。しかしながら、これまでの意思決定に関する理論研究は、ヒトや動物が合理的に行動する主体であることを前提としてきたため、このような好奇心に基づいた非合理的な行動を説明することが困難であった。また、「報酬をとるか、好奇心に従うか」といった心の揺れや葛藤は主観的なもののため、第三者が客観的に数値化することは不可能と考えられてきた。そこで、本研究では、認識と行動選択を統合した自由エネルギー原理に基づき、好奇心の強さを変化させることで、合理的な行動だけでなく非合理的な行動も記述できる意思決定モデル（ReCU モデル）を提案した。さらに、行動データから心理状態の時間変化を読み解く手法「逆自由エネルギー原理法」を開発し、動物の行動データから報酬と好奇心との葛藤を解読した。この手法をラットの行動データに適用した結果、ラットは負の好奇心を持ち、確実な選択肢に固執する保守的な行動を取ることが分かったほか、期待される情報量によって好奇心のレベルが変化することも明らかになった。本手法により、これまで定性的にしか議論できなかった心理的状態を定量的に評価することが可能となるため、「心の揺れ・葛藤」の神経メカニズムの解明に大きく貢献することが期待される。さらには、過度な好奇心の増減は ADHD（注意欠陥多動性障害）や自閉症とも関連すると考えられるため、精神疾患の診断への応用も期待される。

目次

2.背景

2.1 意思決定に関するこれまでの研究

ヒトや動物は、感覚系を通じて外界を認識し、それに応じて体内の状態を変化させるほか、意思決定を行い行動する[1,2]。このような、人や動物が外界を認識し、学習・意思決定する過程を分析した古典的な例としてパブロフの犬が挙げられる。ソビエトの生理学者であるパブロフは、1903 年、鳴き声やベルの音を食事の前に提示することで、犬が音に対して唾液分泌を引き起こす条件反射を形成できることを発見し、刺激と反応の関連性を示す古典的条件づけの理論を提唱した。その後 1938 年に、アメリカの心理学者、バラス・スキナーは、行動により報酬が得られたり、罰が与えられたりすることで、行動の強化や抑制が生じるメカニズムをオペラント学習として定式化した。オペラント学習では、ある行動（反応）が刺激によって、受動的に学習される古典的条件づけとは異なり、ある行動が「強化」という機能によって能動的に学習される。

1998 年、リチャード・S・サットンとアンドリュー・バートは、生体が環境との相互作用を通して学習し、意思決定する過程を強化学習としてモデル化した。強化学習では、古典的条件づけ（将来の報酬と罰を予測）とオペラント学習（将来の報酬と最大化する行動を選択）の両方がモデリングされている。また近年では、カール・フリストンにより、ベイズ推定によって脳が外界を最適に認識するというベイズ脳仮説の基づいた自由エネルギー原理が提唱された[3–5]。自由エネルギー原理は、「能動的推論」[6–8]として知られる外界の認識の不確実性を最小化する能動的な情報探索行動も表現しているほか、期待報酬と好奇心で構成された期待自由エネルギーと呼ばれる行動のスコアを提案し[9–14]、報酬と好奇心の両方を最大化することによって認知と行動を定式化した。なお、期待自由エネルギーにおいて好奇心は情報利得、すなわち、行動を通じた新たな観測により自分の認識がどれほど更新されるのかを表す程度と見なすことができる。これらの強化学習や自由エネルギー原理は、昨今の脳科学・神経科学を主導する理論としての役目を果たしてきた。

2.2 本研究の新規性

ヒトや動物は、感覚系を通じて外界を認識し、それに応じて体内の状態を変化させているが、実際には環境の不確実性に加え、脳の計算能力の限界や意思決定の時間的制約から、最適な意思決定を行うことは難しい[15]。たとえば、我々は報酬が期待できないにもかかわらず、宝くじを引いたり、ギャンブルをしたりと非合理的な行動をとっている。この場合、報酬の期待値の低さと報酬が得られるかどうかという好奇心の間でジレンマが生じており、両者のバランスを制御し行動していると言える。このように、

動物が報酬と好奇心のバランスをどのようにコントロールしているかを理解することは、意思決定プロセスを解明する上で重要である。しかし、報酬と好奇心のバランスを定量的に評価する方法は未だに確立されていない。

また、好奇心の強さゆえに出現する非合理的な行動も存在する[16,17]。例えば、保守的な人は不確実性を避け、予測可能な結果につながるよう行動する。逆に、好奇心の強い人は、報酬よりも環境を知ることに重きをおき、予測不可能な結果につながる行動を選択する傾向にある。好奇心が両極端にふれてしまった保守的すぎる性格と好奇心が強すぎる性格は、それぞれ自閉症スペクトラム障害（ASD）や注意欠陥多動性障害（ADHD）の症状として解釈することができる。実際、これらの精神疾患の患者の性質として新たな情報を忌避することや、逆に強く希求することが知られている[18–25]。合理的な人は、これらの両極端な性格の中間的な性格を示し、曖昧な環境では、環境を効率的に理解するための行動を選択し、環境が明らかになれば、報酬を効率的に獲得するよう行動する。このように、好奇心が行動パターンに大きな影響を与え、動物は文脈依存的に報酬と好奇心のバランスを制御していると考えるのが自然である。

強化学習モデルの特徴は動物が報酬を獲得しようとする行動だけでなく、探索行動も記述する点であり、同モデルにおける探索行動は受動的でランダムな行動選択として表現されている[26]。しかし、実際の動物の行動をみると、好奇心をもつ動物はランダムに行動しているのではなく、環境の不確実性を最小化するように能動的に環境を探索しており、強化学習モデルではこのような好奇心に従った能動的な探索行動は表現できない。また、近年注目を浴びている自由エネルギー原理では、報酬と好奇心の重み付けが常に均等かつ一定であることを仮定しており、報酬と好奇心の重みが時間とともに変化すると考えられる実際の動物行動を扱うことはできない[27]。したがって、強化学習や自由エネルギー原理のような従来の理論は報酬と好奇心の間の葛藤を記述するには限界がある。

好奇心の時間的変動を明らかにすることは、意思決定における報酬と好奇心の葛藤の神経基盤を解明する上で重要である。これまでの自由エネルギーに基づいた研究の多くは、動物の意思決定がベイズ最適であることを前提とした理論構築に終始していた。このため、動物が報酬と好奇心の葛藤によって非合理的に意思決定を行うという発想すらなく、行動データから報酬と好奇心の時間的バランスを読み解く方法は言わずもがな確立されていない。行動データから報酬と好奇心の時間的バランスを読み解く方法を開発すれば、好奇心の時間的変動と神経相関を解析することができ、その結果として、脳が報酬と好奇心のバランスを文脈依存的に制御している様子を明らかにすることができると考えられる。

本研究では、報酬と好奇心の葛藤のダイナミクスを制御するメタパラメータを組み込むことで自由エネルギー原理を拡張した Reward-Curiosity 意思決定（ReCU）モデルを開発した。ReCU モデルでは、報酬に貪欲な行動、高い好奇心による情報探索行動、

不確実性を回避する保守的な行動など、様々な行動パターンを示すことができた。さらに、意思決定における情報処理の内部変数を推定するために、逆自由エネルギー（iFEP）法と呼ばれる機械学習法を開発した。iFEP 法を二者択一課題における行動時系列に適用することで、好奇心の変動、報酬の有無の認識、その確信度といった内部変数の推定に成功した。

3.結果

3.1 報酬と好奇心のジレンマによる意思決定

ヒトや動物は、観測から報酬の有無などの原因を推論することで環境を認識し、自らの推論に基づいて意思決定を行っている。本研究では、同じ報酬でも報酬確率の異なる二つの選択肢のいずれかを選択する二者択一課題において、報酬と好奇心のジレンマに直面したエージェントがどちらか一方の選択肢を選ぶプロセスを ReCU モデルとして定式化した（図 1A)。累積報酬の最大化を目指すのであれば、報酬確率の高い選択肢を選択する必要がある。しかし、動物行動実験では、どちらの選択肢がより報酬と関連しているかを学習しても、専ら最良の選択肢を選択するわけではなく、報酬確率の小さい選択肢を選択することも多い。

このような動物の行動について、本研究では次のようなメカニズムが背後に存在していると考えた。仮説：「動物は、各選択肢の報酬確率が時間とともに変動する可能性があると考え、ある選択肢を選択し続けると、他の選択肢の報酬確率の推定に対する自信が低下する。そのため、エージェントは報酬確率が小さくても曖昧な選択肢を選ぶことで、両選択肢の推定に対する自信を高めている。」こうした仮説に基づき、ReCU モデルを構築した。

3.2 ReCU モデル

ReCU モデルでは、脳内の情報処理を 2 つのプロセスに分けた。最初のプロセスでは、エージェントは各選択肢の報酬確率の認識を更新する（図 1A; プロセス 1)。次に、エージェントは現在の認識と好奇心に基づいて行動を選択する（図 1A；プロセス 2)。エージェントは、これら 2 つのプロセスを繰り返し実行する。

最初のプロセスでは、報酬確率が時間的に潜在的に変動するという仮定のもと、エージェントは逐次ベイズ推定によって報酬確率を推定しているとモデル化した（図 1B)。このモデルでは、エージェントは、行動と結果報酬の観測に応じて、推定分布として表される報酬確率に関する認識を更新する（図 1C)。また、報酬確率に関する認識の更新式を図 1D のように導出した（図 1D；手法参照)。 2 つ目のプロセスでは、エージェントには報酬を最大化する欲求と環境から情報を得る欲求の 2 つの動機が存在し、両者の合計が最大化するよう行動する（図 1E)。この合計は、本研究では「期待純効用」と呼び、以下の式で表す。

$$U_t(a_{t+1}) = E[R_{t+1}] + c_t \cdot E[Info_{t+1}], \quad (1)$$

ここで、第 1 項は現在の認識に基づく次の行動a_{t+1}に対する報酬を表し、第 2 項は新しい観測から得られる情報の期待値を表す(本研究では情報利得と呼ぶ)。c_tは好奇心の強さを表すメタパラメータを示し、期待される情報の期待値を重み付けしている。式(1)の

構造からわかるように、本モデルでは、c_tが時間的に変化することで、非合理的な心の葛藤を表現している（図 1F）。次に、期待純効用に基づきエージェントは行動を選択

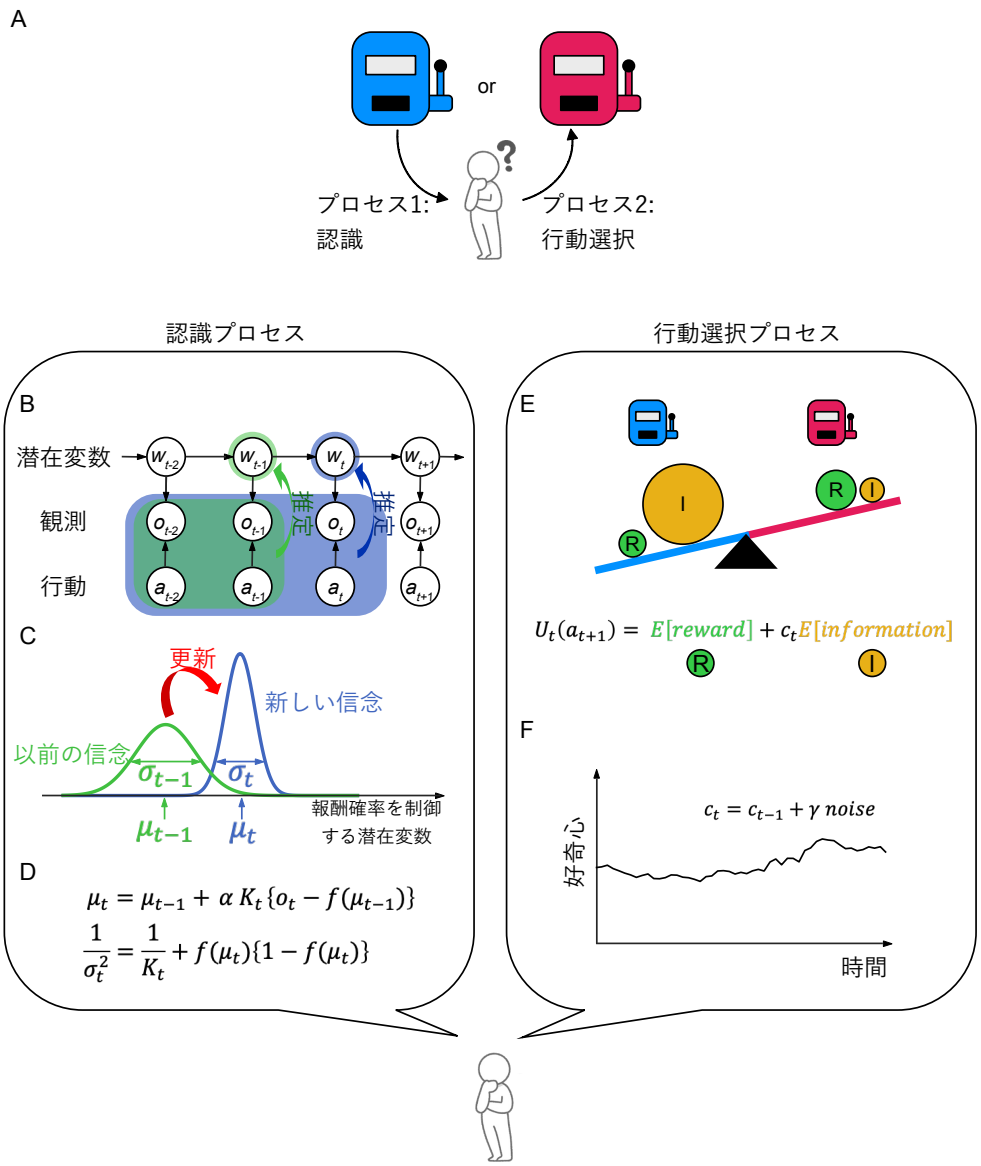


図 1:報酬と好奇心のジレンマを伴う二者択一課題の意思決定モデル

(A)二者択一課題における意思決定の概略。報酬は各選択肢で異なる確率で提供される。エージェントはその確率を知らず、試行錯誤を繰り返しながら、各選択肢の潜在的な報酬確率を推定し、自らの推定に基づいて、どちらの選択肢を選択するかを決定する。(B) 認識プロセスの状態空間モデル。エージェントは、報酬確率を制御する潜在変数が時間変動することで、報酬確率が変化すると認識していると仮定する。(C) 認識の更新の概略。エージェントは潜在的変数を確率分布として認識する。(D)各選択肢の推定分布の平均と分散の更新式。α、K_t、$f(\mu_t)$はそれぞれ学習率、カルマンゲイン、報酬確率の予測値を示す。ただし、両式の第 2 項は、オプションが選択されない場合は消滅する。(E)エージェントによる行動選択過程。エージェントは、式で示されるように、期待報酬と情報利得の加重和を用いて、各行動の期待純効用$U_t(a_{t+1})$を評価する。エージェントは、両者の期待純効用を比較し、より大きな期待純効用を持つ選択肢を好んで選択する。(F)好奇心の時間依存性。好奇心の強さは、c_tの変動によって時間的に変化する。

するが、その際、より高い期待純効用を持つ行動a_{t+1}を選択することを好み、行動はシグモイド関数にしたがって確率的に選択される。

$$P(a_{t+1} = i) = \frac{1}{1 + exp(-\beta \Delta U_t)}, \quad (2)$$

ここで$\Delta U_t = U_t(a_{t+1}) - U_t(\overline{a_{t+1}})$と$\beta$は、行動選択のランダム性を制御する逆温度を表す[28–30]。

本モデルを検証するために、以下の 2 つのケースについて、好奇心$c_t = 1$を一定にしたシミュレーションを行った。最初のケースでは、報酬確率が一定で 2 つの選択肢の間で異なる場合（図 2A）を想定した。このシミュレーションではエージェントは報酬確率の高い選択肢を優先的に選択した（図 2B）。このとき、認識された報酬確率は真実の報酬確率に収束し、エージェントが報酬確率を正確に認識していることが示された（図 2C）。このとき、選択された選択肢の報酬確率に対する自信は観測からの情報によって上昇し、選択されなかった選択肢の報酬確率に対する自信は低下した（図 2D）。同様に、ある選択肢の情報利得は、その選択肢が選択されたときに上昇し選択されなかったときに減少した。したがって、自信のある選択肢の方が期待される情報量は少なくなった（図 2E）。期待報酬は、認識された報酬確率に従った（図 2F）。また、シミュレーション初期において、情報利得は顕著に減少した一方、期待報酬は増加し、両者はやがて交差した。これは、エージェントが最初は情報探索に重きをおいているのに対し、やがて報酬獲得に重きをおくように切り替えていることに相当する（図 3A）。また、これら 2 つの要素は負の相関を持ち、トレードオフの関係にあることがわかる（図 3B）。期待される情報利得と報酬の和である期待純効用は、それぞれの選択の価値を表し（図 2G）、その結果、エージェントは期待純効用の高い選択肢を優先的に選択した。

第 2 のケースとして、報酬確率が時間依存する動的環境を想定した（図 2H）。シミュレーションでは、エージェントは真の報酬確率の変化に応じて報酬確率の認識を適応的に変化させ、より高い推定報酬確率を持つ選択肢を選択した（図 2I、J）。推定に対する自信は各時刻の報酬確率の不確実性に影響され、報酬確率が 1 や 0 ほど近く決定論的な場合は高くなった一方、報酬確率が 0.5 程度と不確実な場合は低くなった（図 2K）。ある選択肢の情報利得は、その選択肢の信頼度と負の相関があり（図 2L）、エージェントは不確実な選択肢に好奇心を抱いていることが示唆された。また、期待報酬は認識された報酬確率と連動して変化した（図 2M）。第 2 のケースにおいても第 1 のケースと同様に、シミュレーションの初期段階において、情報探索と報酬獲得の切り替えが観測された（図 3C）。ただし、第 1 のケースとは対照的に、報酬確率が時間的に変化する環境のため、期待される報酬と情報利得は明確な線形相関を示さなかった（図 3D）。また、期待純効用も期待報酬と同様に変化した（図 2N）。これらのシミュレーションは、本モデルが報酬と好奇心に基づく認知・意思決定のプロセスを表現できることを示す。

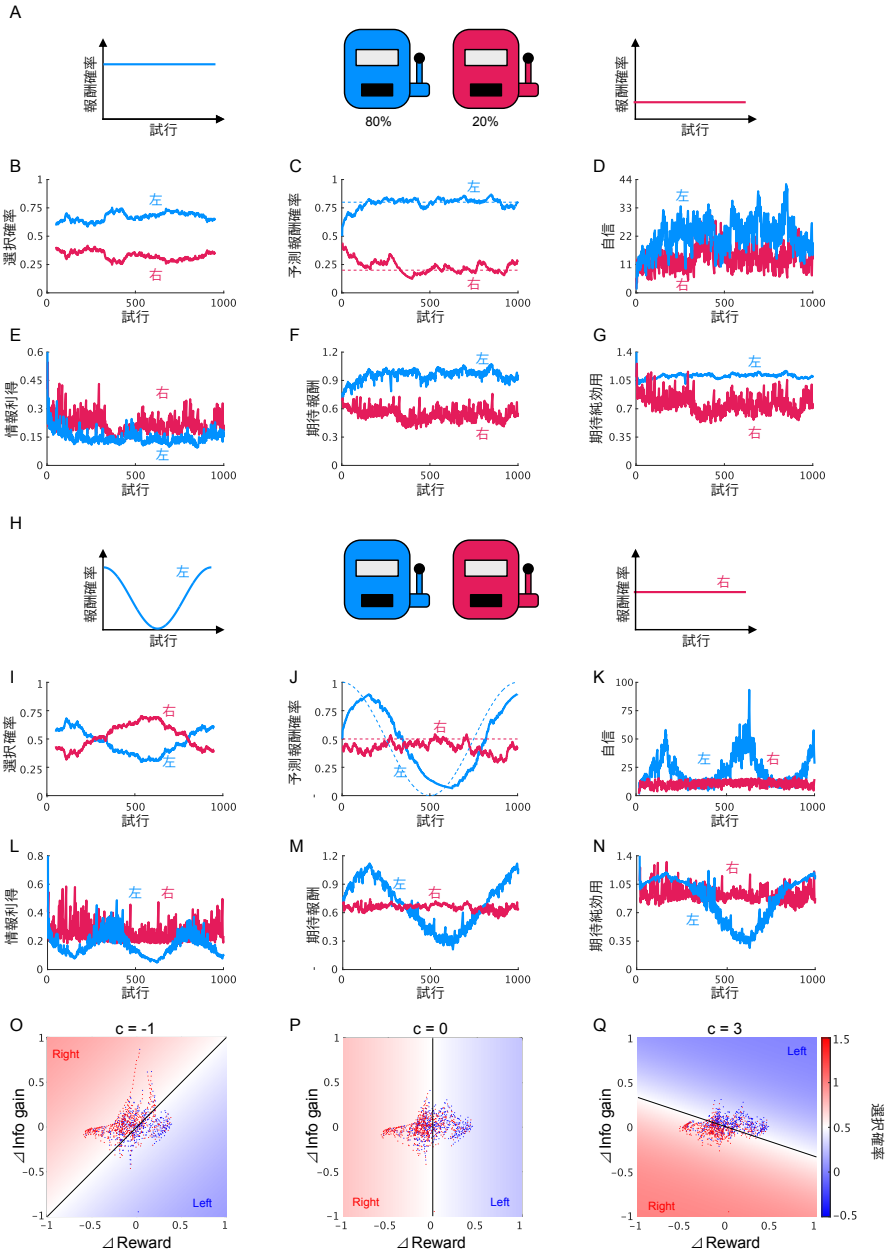


図 2:意思決定モデルのシミュレーション

(A)報酬確率が一定な場合の二者択一課題の概略図。(B)左右の選択肢の選択確率の移動平均（平均範囲は 101 トライアル）。(C)実際の報酬確率とエージェントが推定した報酬確率との比較。(D)左右の選択肢の報酬確率の認識に対する自信。(E)情報利得の推移。(F)期待報酬の推移。(G)期待純効用の推移。(H)報酬確率が変動する場合の二者択一課題の概略図。(I-M) 前述の(B-G)と同様。これらのシミュレーションのパラメータ値は$c = 1$, $P_o = 0.8$, $\alpha = 0.05$, $\beta = 2$ $\sigma_w = 0.63$。(O-Q)期待報酬と情報利得の空間における選択された選択肢の分布。O-Q でそれぞれ好奇心の強さが異なる（(O) $c = -1$, (P) $c = 0$ (Q) $c = 3$）。なお、報酬確率は Ornstein-Uhlenbeck 過程によって遷移する$w_{i,t}$によって生成された。: $w_{i,t} = w_{i,t-1} - 0.01w_{i,t-1} + 0.15\xi_t$、$\xi_t$は標準ガウスノイズを示す。ヒートマップは、空間における行動選択の確率を表す（式（2））。

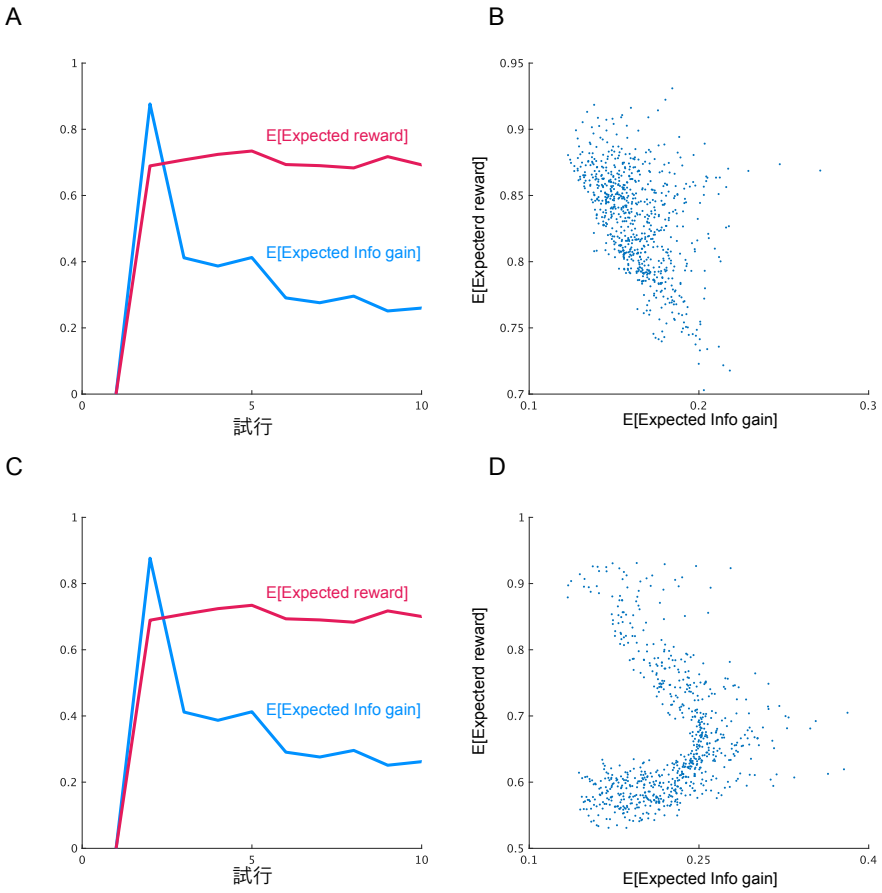


図 3:意思決定モデルのシミュレーション
(A)報酬確率が一定な場合の二者択一課題の概略図。(B)左右の選択肢の選択確率の移動平均（平均範囲は 101 トライアル）。(C)実際の報酬確率とエージェントが推定した報酬確率との比較。(D)左右の選択肢の報酬確率の認識に対する自信。(E)情報利得の推移。(F)期待報酬の推移。(G)期待純効用の推移。(H)報酬確率が変動する場合の二者択一課題の概略図。(I-M) 前述の(B-G)と同様。これらのシミュレーションのパラメータ値は$c = 1$, $P_o = 0.8$, $\alpha = 0.05$, $\beta = 2$ $\sigma_w = 0.63$。(O-Q)期待報酬と情報利得の空間における選択された選択肢の分布。O-Q でそれぞれ好奇心の強さが異なる（(O) $c = -1$, (P) $c = 0$ (Q) $c = 3$）。なお、報酬確率は Ornstein-Uhlenbeck 過程によって遷移する$w_{i,t}$によって生成された。: $w_{i,t} = w_{i,t-1} - 0.01w_{i,t-1} + 0.15\xi_t$、$\xi_t$は標準ガウスノイズを示す。ヒートマップは、空間における行動選択の確率を表す（式（2））。

3.3 受動的な行動と好奇心依存的な行動の判別

本研究では、ヒトや動物は好奇心を持ち、状況依存的に好奇心を変化させて意思決定を行っていると仮定しているが、そもそもヒトや動物が好奇心依存的に行動しているのか確認することは、重要である。そこで、好奇心を持たない受動的な行動と好奇心駆動的な行動を判別する方法を提案した。

期待純効用は次のように書き換えることができる。

$$\Delta U_t = \Delta E[R_{t+1}] + c_t \cdot \Delta E[Info_{t+1}], \quad (3)$$

ここで、第 1 項と第 2 項は、それぞれ 2 つの選択肢の期待報酬と情報利得の差を表し、エージェントは$\Delta E[R_{t+1}]$と$\Delta E[Info_{t+1}]$のバランスに基づいて意思決定する。本研究では$\Delta E[R_{t+1}]$と$\Delta E[Info_{t+1}]$の空間において行動を可視化することで、左右どちらを選択するかは境界線：$E[R_{t+1}] + c_t \cdot \Delta E[Info_{t+1}] = 0$によって分けることができることを明らかにした（図 2O-Q)。すなわち、$c_t = 0$のエージェントは$\Delta E[R_{t+1}]$のみに基づいて選択肢を選択した（図 2P）一方、c_tが 0 以外の値である場合、c_tの正負によって境界が異なる方向に傾いた（図 2O、Q)。これらの結果は、$\Delta E[R_{t+1}]$と$\Delta E[Info_{t+1}]$.の空間における選択行動の分布パターンに基づいて、受動的選択$(c_t = 0)$と好奇心に依存した選択$(c_t \neq 0)$を識別できることを示す。

3.4 好奇心に依存した非合理な行動

次に、好奇心の強さと報酬の希求度合いによって、行動パターンがどのように制御されるかを検討した（図 4)。報酬確率が左：0、右：0.5 であるシナリオ（図 4A）において、好奇心パラメータcと報酬量の制御パラメータP_oを変化させてモデルをシミュレーションした（図 4B；手法参照)。好奇心がなく$(c = 0)$、エージェントが報酬を強く欲している場合（$P_o = 0.99$))、エージェントは報酬確率の高い右の選択肢を優先した（図 4C-point a)。また、好奇心が強く（$c = 10$)、報酬を欲しない（$P_o = 0.5$）場合でも、より報酬確率の高い右の選択肢を優先した（図 4C-(point b))。この行動は一見合理的に見えるが、エージェントは報酬を求めたのではなく、好奇心に基づき情報を求めた結果、不確実な選択肢を選好したのである。好奇心が負の場合（$c = -10$)、エージェントは最初の選択によって 2 つの選択肢のどちらかを連続的に選択した（図 4C-(point c))。このような固執的な行動は ASD 患者が新しい情報を避け、同じ選択を繰り返す保守的な性質を表す[18–23]。

報酬確率を左が 0.5、右が 1 とした別のシナリオでは、非自明な結果が得られた(Fig.3D-F)。先のシナリオと同様に、報酬に対する欲求が強いエージェント（$P_o = 0.99$、ほぼ 1 に等しい）は、報酬確率の高い右の選択肢を優先した（図 4F-(point a))。一方、報酬に対する欲求がなく$(P_o = 0.5)$、好奇心が強い$(c = 10)$エージェントは、報酬確率の低い左の選択肢を好んだ(図 4F-(point b))。この非合理的な行動は、報酬を求めず好奇心を満たすことに注力した結果であり、ADHD 患者が新しい情報を非合理的に探索することを想起させる[24,25,31]。さらに、先のシナリオ（図 4C-(c)）で見られたように、好奇心がマイナス（$c = -10$）のエージェントは、報酬の欲求とは無関係に、保守的な選択を行った（図 4F-(c))。これらの結果から、行動パターンは報酬と好奇心の度合いに大きく依存することが明らかになった。

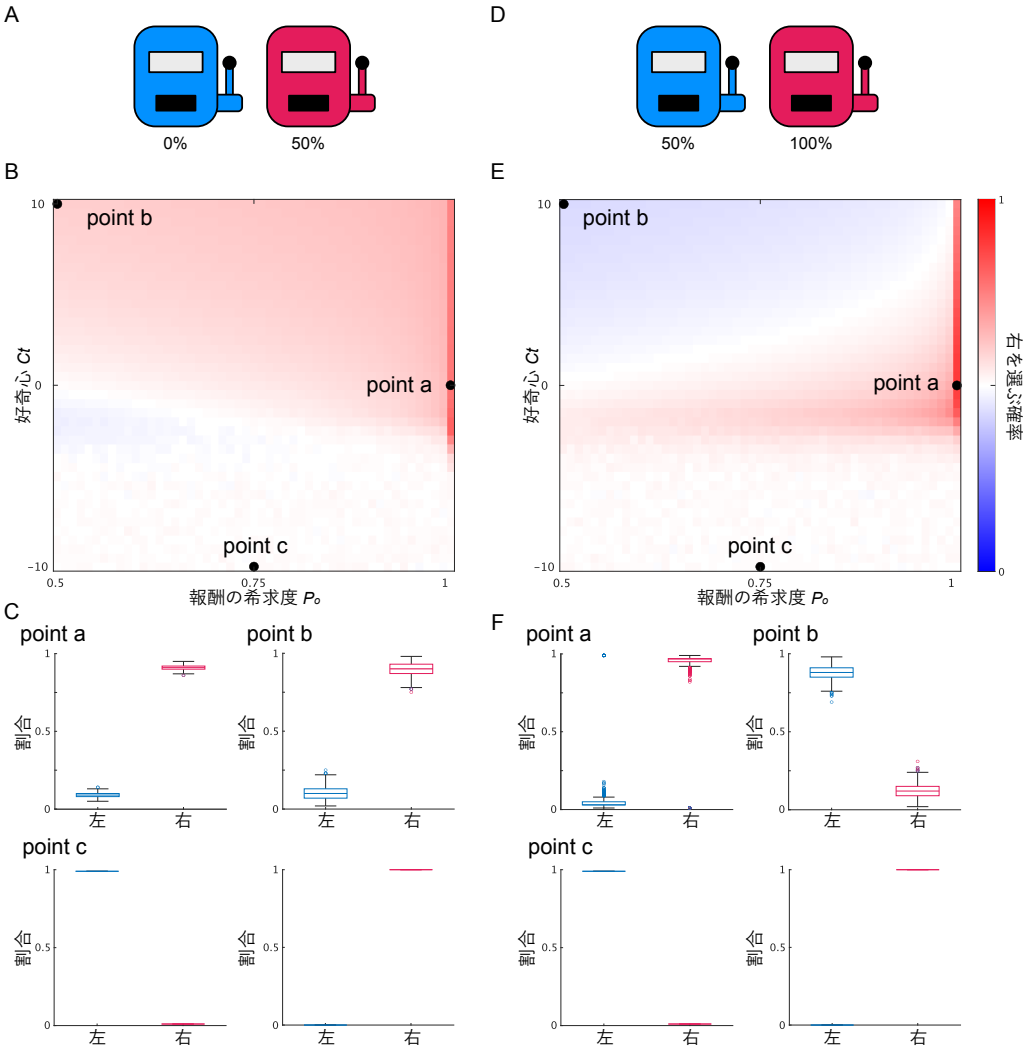


図 4：好奇心に依存した非合理的な行動

(A)異なる一定の報酬確率（左が 0%、右が 50%）を持つ二者択一課題。(B)好奇心と報酬の強さのパラメータを変化させ、右の選択肢を選択する確率をヒートマップで示した。ヒートマップで表した、右の選択肢の選択確率は、好奇心と報酬の各セットについて 1,000 回のシミュレーションを行い求めたものである。(C) 黒点で示した 3 つの代表的な条件（$c = 0$、$P_o = 1$ [poomt a]、$c = 10, P_o = 0.5$ [point b], and $c = -10, P_o = 0.75$ [point c]）における左右の選択比率を箱ひげ図で示した。これらの箱ひげ図において、中央の線は中央値、箱の両端は四分位数、上下のひげは外れ値を除いた最大値と最小値を表す。point c では、エージェントは右か左のどちらかの選択肢を優位に選択する。point c の箱ひげ図は、データ点が密集しすぎて潰れているように見える。 (D) 異なる一定の報酬確率（左が 50%、右が 100%）を持つ二者択一課題。(E)・(F)前述の(B)・(C)と同じ。

3.5 iFEP：ベイズ推定による報酬と好奇心の葛藤の推定

上記のケースでは、報酬と好奇心のバランスが一定であることを想定していたが、実際には私たちの気持ちは状況に依存した形で揺れ動く。このため、報酬と好奇心の葛藤の時間的な揺れを読み解くメタベイズ推論[32–36]が、神経科学や心理学の観点から重要である。本研究では、行動データから好奇心メタパラメータを含む内部状態の時間

的ダイナミクスを定量的に読み解く、逆自由エネルギー原理法（iFEP）と呼ばれる機械学習法を開発した。

iFEP の開発には、エージェントからエージェントの観測者、つまり動物から我々への視点の切り替えが必要である。本研究において ReCU モデルのシミュレーションでは、エージェント視点の状態空間モデル（以下、SSM）を想定し、エージェントが報酬確率を逐次認識する様子を表現した（図 1B、図 5A、B）が、iFEP では逆に、エージェントの内部状態、例えば、好奇心の強さc_t、認識度$\mu_{i,t}$、その確信度$p_{i,t}$（すなわち、推定分布の分散の逆数）を推定するために、観測者視点からの状態空間モデル（観測者-SSM）を組む必要がある（図 5C）。観測者-SSM では、好奇心の強さは連続的に時間変化し、報酬確率の認識は図 1C の式によって更新されるとした。また、エージェントの行動は、式(2)のように、好奇心、認識、自信の強さに応じて生成されると仮定した。ただし、観測者はエージェントの行動と報酬の有無しか観測できない。iFEP では、観測者-SSM に基づき、観測 x からエージェントの潜在的な内部状態 z を以下のようにベイズ推定する。

$$P(z_{1:T}|x_{1:T}) \propto P(x_{1:T}|z_{1:T})P(z_{1:T}), \tag{4}$$
$$z_{1:T} = \{\mu_{i,1:T}, p_{i,1:T}, c_{1:T}\}, x_{1:T} = \{a_{1:T}, o_{1:T}\},$$

ここで、$1{:}T$はステップ 1 からTを意味する。このベイズ推定において、事後分布$P(z_{1:T}|x_{1:T})$は、不確実性を伴う観測値$x_{1:T}$が与えられたときの推定$z_{1:T}$に対する観測者の認識を表す。事前分布$P(z_{1:T})$は、エージェントの認識を表す。また、好奇心メタパラメータ c は以下のようにランダムウォークするとする。

$$c_t = c_{t-1} + \epsilon\zeta_t, \tag{5}$$

ここで、ζ_tはホワイトノイズを、ϵはそのノイズ強度を示す。尤度$P(x_{1:T}|z_{1:T})$は、同じく ReCU モデルに従う$z_{1:T}$を所与とした際に、$x_{1:T}$が観測される確率を表す。この iFEP は、粒子フィルターとカルマンバックワードアルゴリズムを用いて行われた（「手法」参照）。

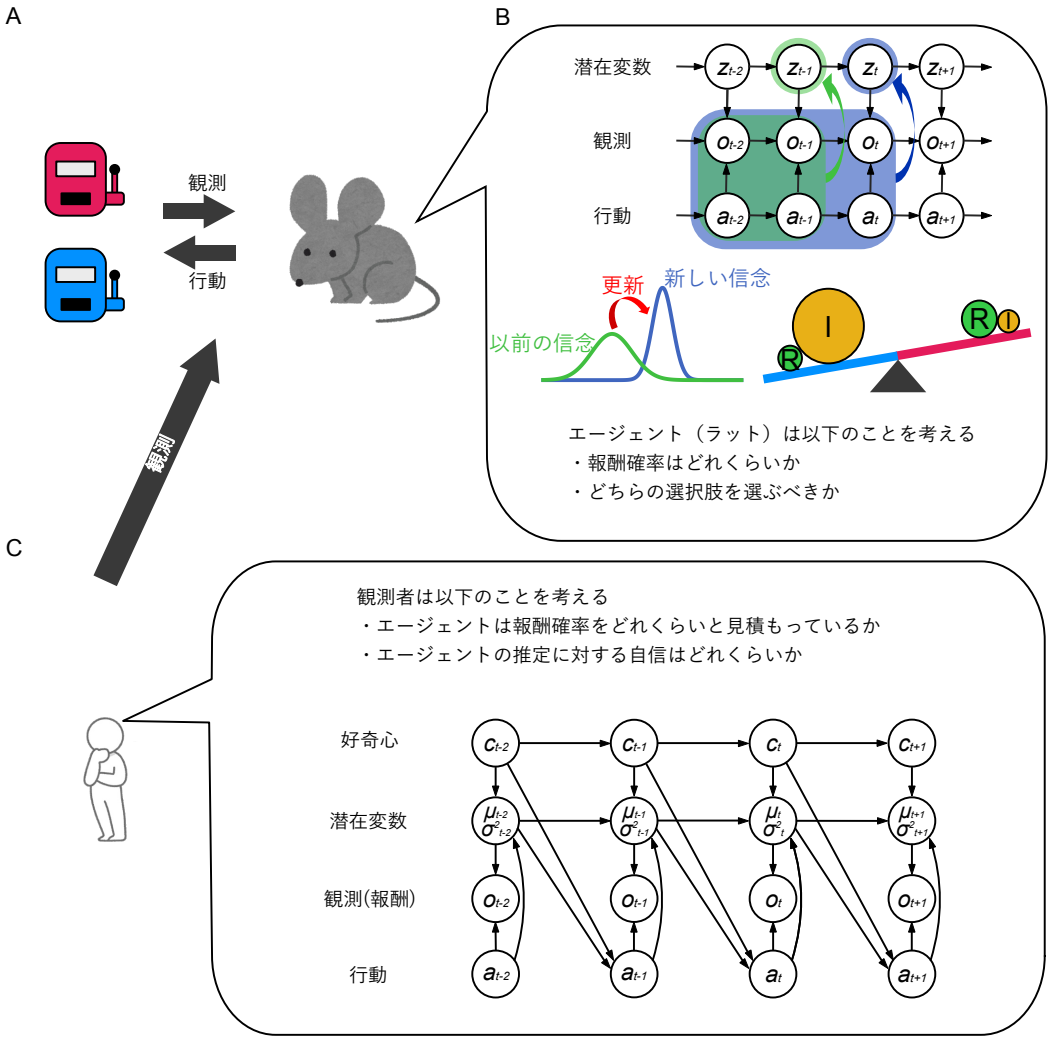


図 5 : iFEP の概念図

(A)二者択一課題を実行するエージェント。(B)エージェントの意思決定に関する観測者の仮定。エージェントは図 1(図 2B に同じ)のような意思決定モデルに従うと仮定した。(C)観測者視点の状態空間モデル。観測者にとって、エージェントの行動と報酬の有無が観測可能であるのに対し、エージェントの報酬と好奇心の葛藤、認識された報酬確率、およびそれらの不確実性は、時間的に変化する潜在的な変数である。こうしたもとで、観測者はエージェントの潜在的な内部状態を推定する。

3.6 人工データによる iFEP の検証

ReCU モデルで生成した人工データに iFEP 法を適用し、その妥当性を検証した。報酬確率が時間的に変化する二者択一課題において、好奇心が一定でないモデルエージェントの行動をシミュレーションした。そして、iFEP がモデルエージェントの内部状態、すなわち好奇心、認識、自信の強さの値を推定することを実証した（図 6）。また、その推定性能は ε の値に対してロバストであることを確認した（図 7）。したがって、iFEP は意思決定プロセスとそれに伴う報酬と好奇心の葛藤の時間的揺らぎを明らかにするために有効である。

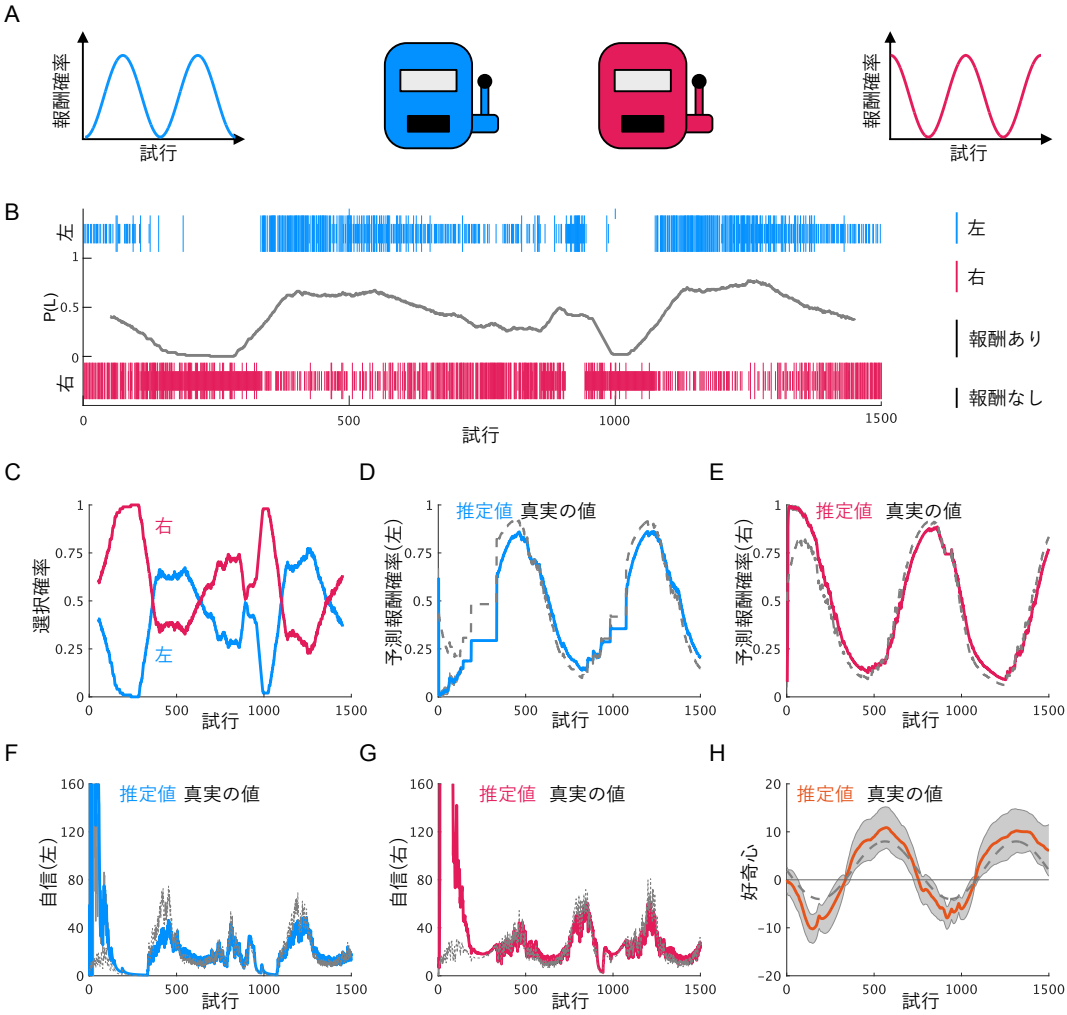


図 6：iFEP による模擬エージェントの内部状態の推定

(A) 好奇心が時間的に変化するエージェントによる報酬確率が時間的に変化する二者択一課題。(B) 模擬エージェントによる二者択一課題のシミュレーション。縦線は左と右の選択肢をそれぞれ選択したことを示す。時系列折れ線グラフは、左の選択肢の選択確率の移動平均（平均を取る範囲は 101 トライアル）。(C) 左と右の選択肢の選択確率の移動平均。(D-H) 左(D)と右(E)の選択肢に対しエージェントが認識した報酬確率、左(F)と右(G)の選択肢に対しエージェントが認識した報酬確率に対する自信、エージェントの好奇心の推定値(H)。本シミュレーションでは、パラメータ値としてP_o=0.8、α=0.05、β=2、σ=0.2 を用いた。また、粒子フィルタの粒子数は 100,000 個で行った。影は標準偏差を表す(D-H)。

3.7 iFEP によるラットの報酬と好奇心の葛藤の推定

報酬確率を時間的に変化させた二者択一課題 [37] の実際のラットの行動データに iFEP を適用した（図 8A）。この実験では、報酬確率を離散的に変化させると、ラットは徐々に報酬確率の高い選択肢を選択し（図 8B）、ラットが報酬確率の認識を逐次更新していることが示唆された。本研究では、iFEP を用いることで、これらのラットの行動データから、ラットの内部状態、すなわち好奇心の強さ、認識した報酬確率とその推

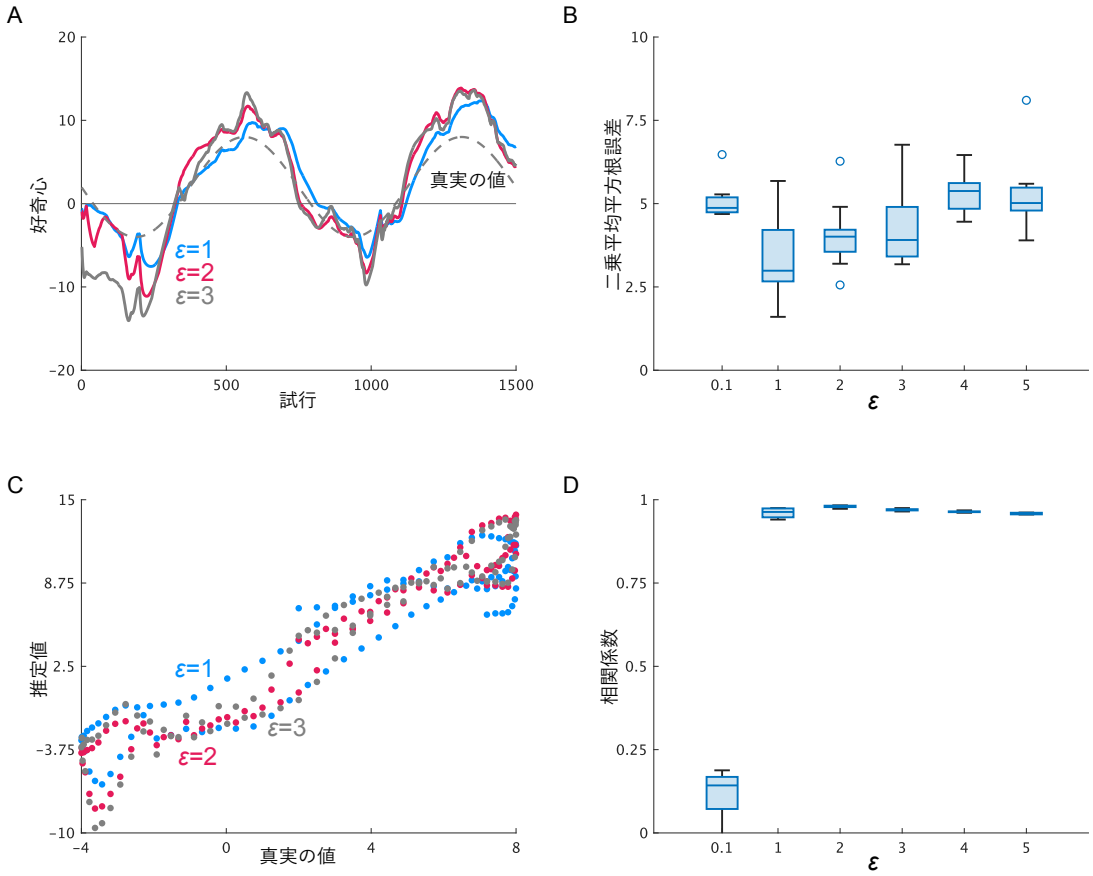


図 7:ϵに依存する iFEP の性能

(A)異なるϵでの好奇心の推定。ReCU モデルで生成した人工データから iFEP で好奇心を推定した。(B) 様々なϵに依存する真実の好奇心の値と推定値との間の二乗誤差。推定は異なるランダムシードで 10 回繰り返した。(C) 異なるϵにおける真実の好奇心の値と推定値の関係。このプロットでは、500 回目から 1,500 回目の試行から 10 間隔ごとに試行を抽出した（各ϵについてn = 101）。(D) 様々なϵに依存する真の好奇心の値と推定した好奇心の値の間の相関係数。推定は異なるランダムシードで 10 回繰り返された。すべての推定(A-D)において、粒子フィルタの粒子数は 100,000 である。箱ひげ図（B, D）において、中央の線は中央値、箱の両端は四分位数、上下のひげは外れ値を除いた最大値と最小値を示している。

定への自信を推定した（図 8C-E)。その結果、ラットは真の報酬確率を完全には認識していないが、報酬確率の増減を認識できることが明らかになった（図 8D、E)。また、自信は選択すれば増加し、選択しなければ減少することがわかった（図 8F、G)。iFEP では、ラットは環境が変化すると認知しているのか、不変と認知しているのかを調べることもできる。行動データから、報酬確率の制御変数w_tの精度は$p_w = 1.785$ (i.e., $\sigma_w^2 = 0.560$)であると推定された。これは、報酬確率の制御潜在変数が、試行とともに標準偏差が増加するランダムウォークを示すことを意味する。これをもとに報酬確率を計算すると、わずか 10 試行で 0.5 から 0.5±0.4 まで大きく変化する（図 9)。このことから、ラットは変動する環境を想定しているほか、認識を忘却し認識に対す自信も喪失することが示唆された。

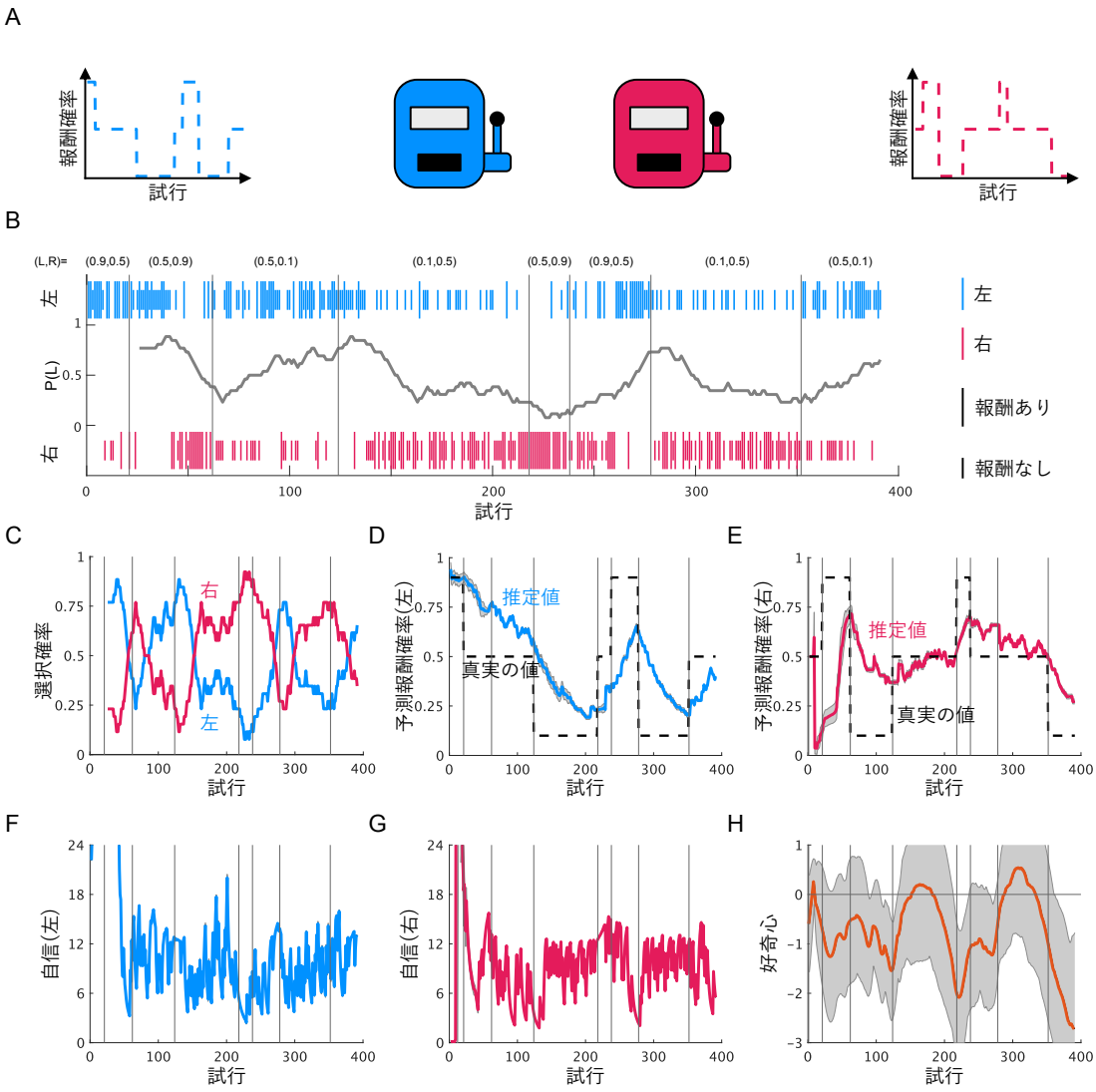


図 8: iFEP によるラットの内部状態の推定

(A)報酬確率が時間的に切り替わる二者択一課題。(B)実際のラットの行動データ。縦線はラットが左右のいずれの選択肢を選択したかを示す。時系列折れ線グラフは、左の選択肢の選択確率の後方移動平均(平均の範囲は 25 トライアル)である。(C)左と右の選択肢の選択確率の移動平均。(D, E)左(D)と右(E)の選択肢におけるエージェントが認識した報酬確率のラット行動データによる推定値。(F-H)ラットの行動データから推測された、左(F)と右(G)の選択肢の報酬確率を認識したエージェントの自信と、エージェントの好奇心の推定値(H)。推定されたパラメータ値は、$\alpha = 0.058$, $\beta = 6.991$, and $\sigma_w^2 = 0.560$。粒子フィルタの粒子数は 100,000 である。影は全粒子の標準偏差を表す（D-H)。

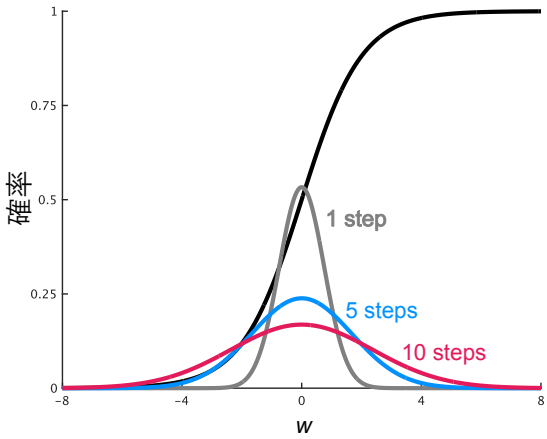


図 9: 観測を伴わない場合、経時的に認識が曖昧になることのシミュレーションによる実証
ReCU モデルではエージェントがwのランダムウォークを想定していることから、一方の選択肢を選択すると、もう一方の選択肢の報酬確率の認識が曖昧になる。図のシグモイド関数は、wに依存する報酬確率を表し、3 つのガウス分布は、w=0 から始まる 1,5,10 回のそれぞれの試行後の報酬確率の認識のブレを表している。iFEP によるラットの行動データからσ^2=0.560 と推定され、わずか 10 試行で推定報酬確率は 0.5 から 0.5±0.4 と有意に変化することが確認できる。

3.8 好奇心の制御メカニズム

iFEP により推定されたラットの好奇心に注目すると、興味深いことに、ラットの好奇心はほぼ全ての試行で負と推定された（図 8H）。つまり、実験で用いたラットは報酬確率の認識があいまいな選択肢を避け、認識が明確な選択肢を好んで選択する、保守的な性格を有するということが分かった。この保守的な行動は、動物が報酬を安定して得ようとするためのものと解釈することができる。負の好奇心を有することを検証するために、図 2O～Q と同様の方法で、$\Delta E[R_{t+1}]$と$\Delta E[Info_{t+1}]$の空間において選択行動を視覚化した（図 10A～C）。推定したc_tが負の場合、$\Delta E[R_{t+1}]$が正で$\Delta E[Info_{t+1}]$が負の領域でラットは左を優位に選択し（$c_t < -1.1$：図 10A，$-1.1 \leq c_t < -0.7$：図 10B）、明らかにラットは正の報酬に対する欲求と負の好奇心を持っていることが分かった。また、推定c_tが 0 に近い場合、ラットは$\Delta E[R_{t+1}]$に基づいて、$\Delta E[Info_{t+1}]$とは無関係に両方の行動を選択した（$-0.7 \leq c_t$：図 10C）。さらに、負の好奇心をもつことは統計的にも検証された（図 11、$p < 0.01$）。

さらに、報酬確率が突然変化すると、好奇心が増加することが明らかになった（図 8H）。この好奇心の変化は、ラットがルール変更を認識し、情報を求める度合いを適応的に制御していると解釈できる。これを受けて、本研究では、ラットが認知した環境に対する情報によって、好奇心をどう制御するかを検討した。まず、好奇心の推定値と期待される情報利得の関連に着目したところ、両者には相関は見られなかった（図

10D, E)。もっとも、好奇心の推定値の時間微分値と情報利得の合計には高い相関がみられた（図 10F, G)。これらの結果は、報酬確率の不確実性が高くなると、好奇心のレベルを積極的に上昇させることを示唆する。

$$\frac{dc}{dt} \propto \sum_i E[info_i]. \tag{6}$$

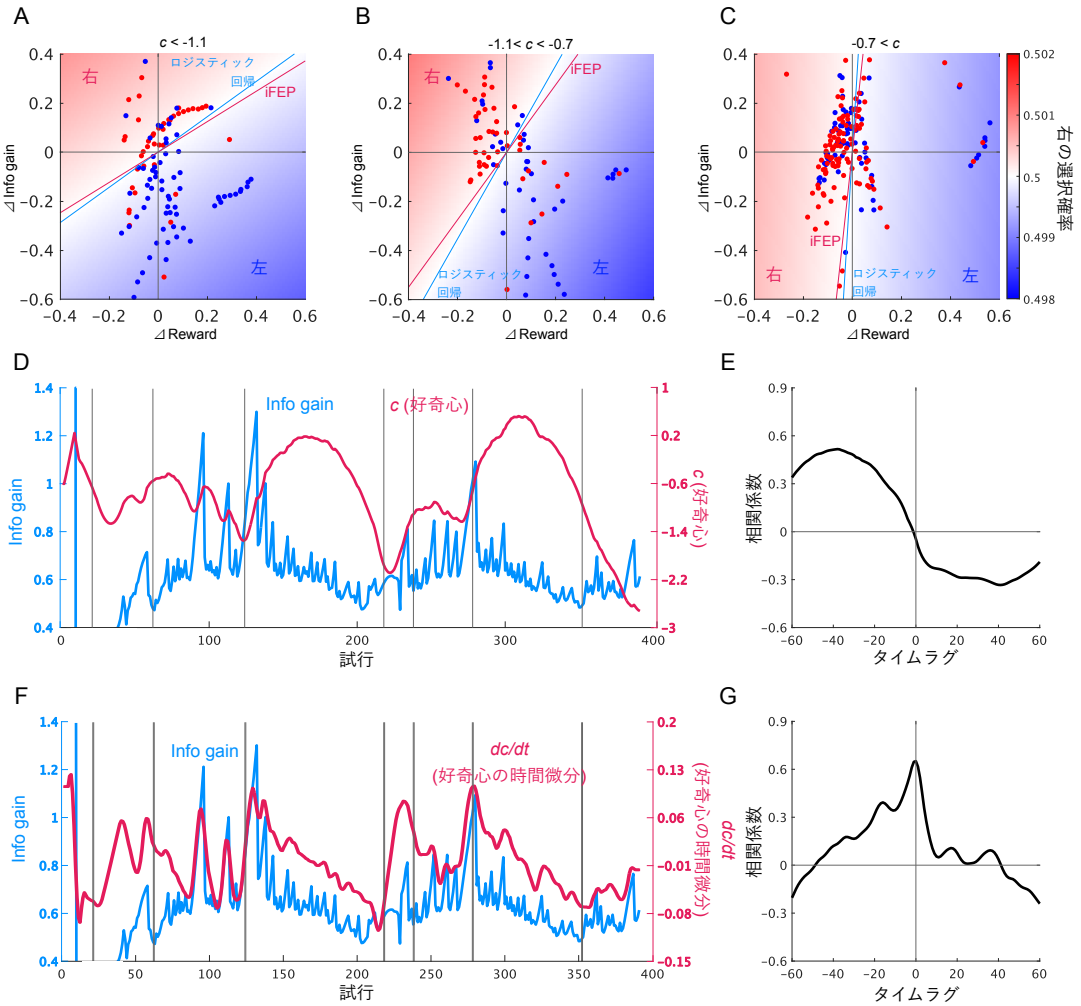
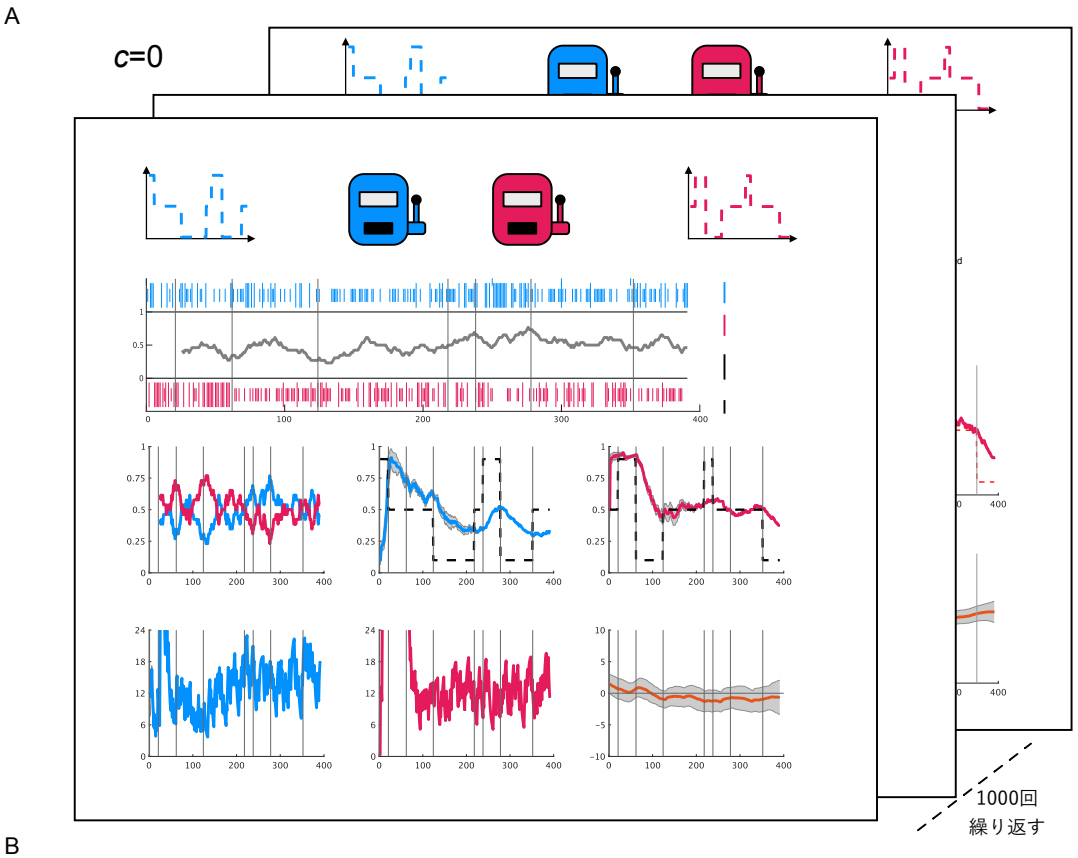


図 10: 負の好奇心とそのダイナミクス

(A-C)期待報酬と情報利得の左右差の空間における選択の分布。すべての行動は、好奇心の推定値が（A）負に大きい場合（$c \leq -1.1,\ n = 110$)、（B）負に中程度の場合（$-1.1 < c \leq -0.7,\ n = 93$)、（C）0 に近い場合（$-0.7 < c,\ n = 187$）に分けられる。2 本の直線は、ロジスティック回帰（$P(a_t = 1) = f(w_R \Delta E[R_{t+1}] + w_I \Delta E[Info_{t+1}])$によりこの散布データから推定した$w_R$と$w_I$を用いた場合と、推定した好奇心の平均値である$w_R = 1$、$w_I = \sum c_t / N$を用いた場合の判別境界を示す。ヒートマップは、推定されたw_Rとw_Iに基づく左の選択肢を選択する確率を表す。(D, E)両選択肢の好奇心の強さと情報利得の総和の時系列(D)、およびそれらの相互相関(E)。(F, G)好奇心の時間微分と両選択肢の情報利得の和の時系列(F)、およびそれらの相互相関(G)。時間微分は、7 試行分の時間窓の中で線形回帰により計算された。



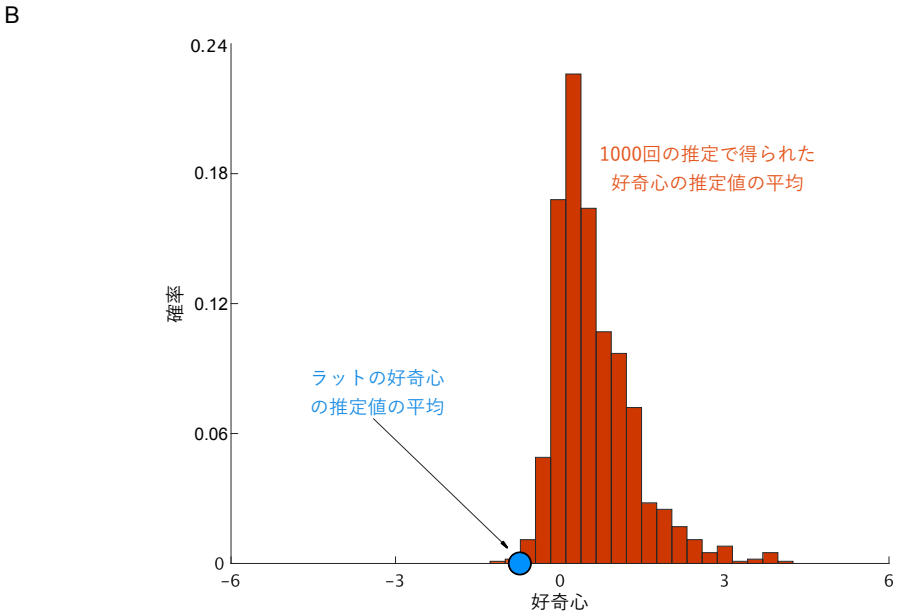


図 11:ラットの好奇心が負であることのモンテカルロ統計検定による検証

（A）ラットに好奇心がない（c_t=0）という帰無仮説のもと、図 8 と同じ実験条件でのモデルシミュレーションを 1,000 回繰り返し、それぞれについて iFEP を用いて好奇心を推定した。推定で用いた粒子フィルタの粒子数は 100,000 個である。(B) 1,000 回のシミュレーションで推定された好奇心の時間平均の無変化分布。ヌル分布と比較して、実際のラットの行動から推定された好奇心の時間平均値は、有意水準より左側に位置していた（p=0.003)。

3.9 代替モデルの評価

最後に、ReCU モデルの妥当性を確認するために、ラットの行動データに基づく他の意思決定モデルとの比較を行った。期待純効用の異なる定式化として、時間依存的な報酬欲求を導入し以下のようなモデルを考えた。

$$U_t(a_{t+1}) = d_t \cdot E[R_{t+1}] + E[Info_{t+1}], \tag{7}$$

ここで、d_tは主観的報酬を記述するメタパラメータを表す（詳細は方法を参照）。この代替モデルを用いて、ラットの行動データから iFEP によりd_tの時系列を推定した。その結果、主観的報酬メタパラメータd_tの推定値は、ラットが報酬確率の急激な変化に遭遇すると動的に変化し、時にはゼロに近くなることがわかった（図 12）。これは、ラットが突然報酬を必要としなくなったことを示し、実験課題前に飢餓状態にして餌を得る動機付けをした動物にとっては不自然なことである。したがって、代替モデルはラットの行動を記述するのに適していないと言える。

また、意思決定タスクのモデルとして広く用いられている Q 学習モデル（強化学習の 1 種）もモデルとして考えられる。本研究では、比較のために先行研究 [37–39] に従い、行動選択のランダム性（好奇心のようなもの）を制御する時間依存の逆温度B_tを導入し、B_tの時系列を推定した（手法参照）。推定の結果、報酬確率が突然変化したときにB_tは減少し（図 13）、ラットは環境ルールの変化に応じてランダムに行動を選択する傾向があることがわかった。このように、環境ルールの変化に応じて逆温度B_tが変化することは合理的であると考えられる。そこで、B_tも ReCU モデルの好奇心パラメータc_tのように、認識の不確実性によって制御されていると仮定し、B_tと情報利得を比較した。ただし、前者は Q 学習モデルに基づいて推定し、後者は ReCU モデルの iFEP によって推定した。その結果、両者は以下のように正の相関があることが明らかになった（図 13）。

$$B_t \propto \sum_i E\left[info_{t,i}\right]. \tag{8}$$

この Q 学習モデルにより推定されたB_tが、ReCU モデルにより推定された情報利得と相関を持つという結果は、ReCU モデルが Q 学習モデルを包含していることを示すとともに、ReCU モデルの説明能力の高さを示唆する。

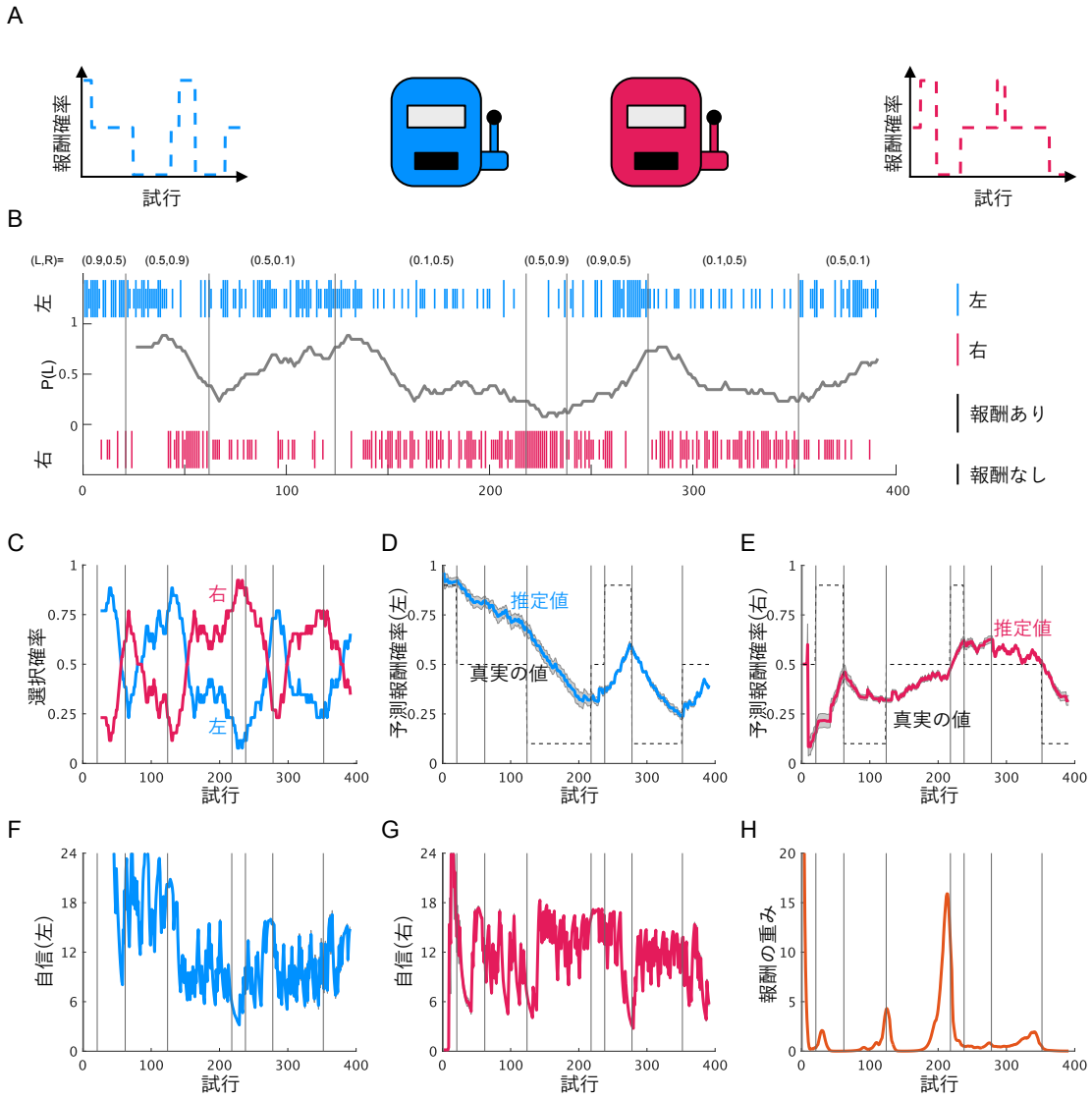


図 12 iFEP によるラットの主観的報酬の推定

(A-G) 図 8 と同様。(H)代替期待純効用($U(a_{t+1}) = d_t \cdot E[R_{t+1}] + E[Info_{t+1}]$)を用いて推定した、報酬パラメータの推移。なお、推定されたパラメータ値は、α=0.051, β =3.585, σ^2 =0.360 であった。粒子フィルターで用いた粒子数は 100,000 個とした。影は、全粒子の標準偏差を表す（D-G)。

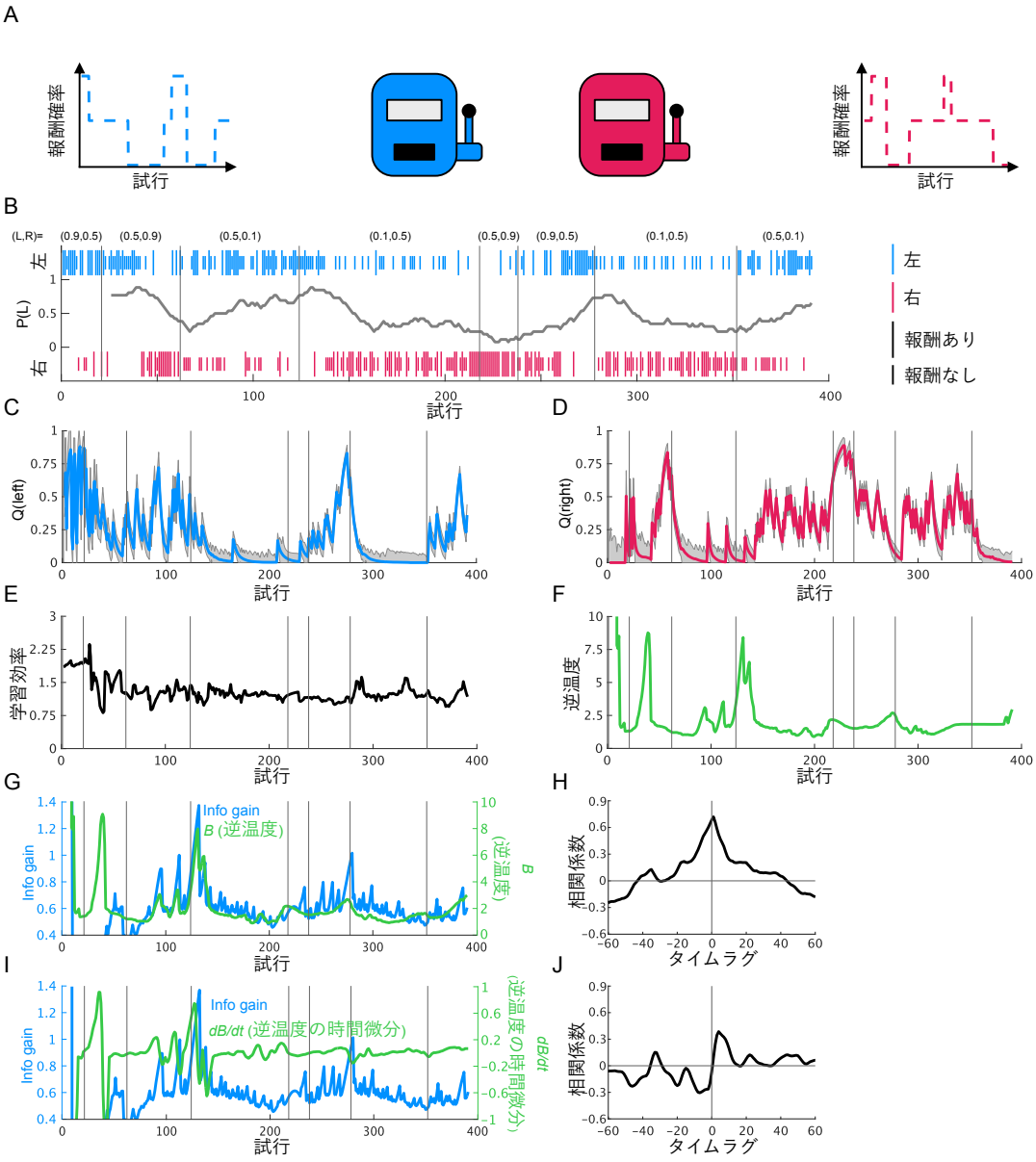


図 13 Q 学習モデルを用いたラットの内部状態の推定

(A, B) 図 8 と同じ。 (C, D) ラット行動データによる左(C)と右(D)の選択肢の行動価値関数Qの推定値。(E, F)ラットの行動データを用いたエージェントの学習効率 α(E)と逆温度 β(F)の推定値。(G, H)逆温度の強度と両選択肢の情報利得の和の時系列(G)、およびそれらの相互相関(H)。(I, J)逆温度の時間微分の時系列と、両者の情報利得の和(I)、およびその相互相関(J)。時間微分は 7 試行分の時間窓で線形回帰により計算した。推定で用いた粒子フィルタの粒子数は 100,000 個である。

4.議論

これまでの意思決定に関する理論研究では、合理性や最適性を前提とした意思決定が議論されており、動物特有の非合理的な行動を扱うことができなかった。これに対し、我々は報酬と好奇心の間の心の葛藤を表現する ReCU モデルを提案したほか、時系列行動データから報酬と好奇心の葛藤を読み解く iFEP 法を開発した。iFEP 法をラットの行動データに適用したところ、ラットの好奇心の値はほとんどの試行で「負」であると推定された。また、推定した好奇心と認識を比較したところ、ラットは報酬確率の認識があいまいになると、好奇心のレベルを積極的に上昇させることが明らかになった。このように、動物が現在の認識と不確実性の度合いに応じて、好奇心を適応的に制御していることを定量的に示した報告は、これまでに例がなく、重要な成果と言える。

本研究のアプローチは、従来のモデルと比較して、3 つの特徴がある。第 1 に、意思決定における非合理性を仮定した点である。期待自由エネルギーはもともと、情報利得と期待報酬の和で表され、両者は等しく重み付けされ、ベイズ最適に導かれていた。しかし、ヒトを含む動物は、限られた時間の中、脳という限られた計算資源のもとで、意思決定を行うため、ときには非合理的な意思決定も行ってしまう。このような性質を捉えるべく、本研究では好奇心を制御するメタパラメータを導入することで、非合理的な意思決定を定式化した。第 2 の特徴は、状況依存的に変化する好奇心を扱ったことである。日常的にヒトや動物の好奇心は時間的に変化し、報酬と好奇心の間のジレンマに直面している。本研究で提案した ReCU モデルでは、そのように揺れ動く好奇心やそれに伴う葛藤を表現できる。第 3 の特徴は、好奇心の強さを含む心理状態を動物の観測者の視点から定量的に解読する iFEP 法を用いて、逆問題に取り組んだ点である。本手法は、好奇心旺盛、保守的といった個体固有の特性だけでなく、状況に依存した時間的な心の揺らぎも評価することができる。

3.9 節で取り上げた Q 学習モデルのような従来の強化学習モデルでは、探索の度合いを逆温度パラメータで表すのが一般的であり、行動選択のランダム性は得られるが、ReCU モデルで表現したような能動的に情報を探索する行動は表現できなかった。これまで、Q 学習モデルをベースに行動データから内部状態を解読するという研究はなされてきたが、そこで推定される好奇心はあくまで「ランダムさ」であり、環境を知るために能動的に情報を獲得するという好奇心とは異なる [37,39–41]。一方、Schwartenbeck らは、自由エネルギー原理に基づき、二者択一課題における意思決定行動をモデル化した [27] が、彼らは好奇心の強さをベイズの最適性が満たされる値（$c_t = 1$）で固定されていると仮定し、本研究で推定したような時間的変動には触れていない。Ortega と Braun は、非合理的な意思決定を記述する自由エネルギー原理を定式化した [42]。彼らの定式化は微視的な熱力学に基づいており、逆温度パラメータがこの非合理性を制御するものであるが、行動データから内部状態を推定するような逆問題には取り組んでいない。これらの従来の研究との比較で分かるように、非合理性と時間変動性を含んだ形で行動データから意思決定に関わる要素を推定した本研究は、

先駆的であると言える。

ReCU モデルにおいても Q 学習モデルにおいても、行動選択はシグモイド関数で定式化されている。したがって、どちらのモデルも時間的に変化する変数をフィッティングすることで、同程度に尤度を上げることができ（詳しくは方法参照）、ReCU モデルと Q 学習モデルのどちらが動物の意思決定に使われているかを尤度によって判別することは本質的に困難である。しかし、iFEP は、時間依存のメタパラメータが認識とその自信によってどのように制御されているかを知る手がかりとなる一方、Q 学習モデルでは環境認識の更新がない動物の行動を説明できず、iFEP は心の葛藤がどのように制御されているかを理解する上で、Q 学習モデルよりも高い一般化能力を有していると言える。

心の葛藤を含む心理状態を制御している神経基盤は何なのか？この問いは、神経活動と動物の行動を比較するだけでは解決できない。なぜなら、心理状態は行動そのものから反映されるのではなく、一連の行動の背後にあるはずだからである。したがって、心理状態を定義し、好奇心、自信、報酬予測誤差などの潜在的な心理状態を動物行動から推定し、推定された心理状態と神経活動を比較することが重要である。この点、動物の行動データから内部状態を推定する iFEP は心理状態を対象とした神経科学の今後の発展に寄与するものであるといえる。

最後に、iFEP の医療における将来的な展望を述べる。一般に、うつをはじめとする精神疾患の診断は大部分を医師による問診に頼っているが、iFEP 法を用いることで、患者の行動データから心理状態を定量的に推定することができるようになる。例えば、ひきこもり患者は、ReCU モデルでは好奇心の値がマイナスになるはずである。このように、iFEP 法は神経科学の基礎研究分野だけでなく、精神医学領域においても強力なツールになり得る。

5.手法

5.1 報酬量

二者択一課題では、報酬は「あり」・「なし」の 2 択だが、その強さはエージェント自身の感覚に依存し、以下のように記述できる。

$$R = 0 \; or \; \ln \frac{P_o}{1 - P_o} \tag{9}$$

ここで、P_oはエージェントがどれだけ報酬を欲しているかを表し、エージェントが感じている報酬強度を制御する(図 14)。自由エネルギー原理の提唱者であるフリストンは、報酬は$\ln P_o$と$\ln(1 - P_o)$として表現したが [10,28]、フリストンによる定式化では、報酬が負の値を取ってしまうため、本研究では式(9)のように定式化した。もっとも、本質的にはフリストンによる定式化も本研究での定式化も意味するところは同じである。

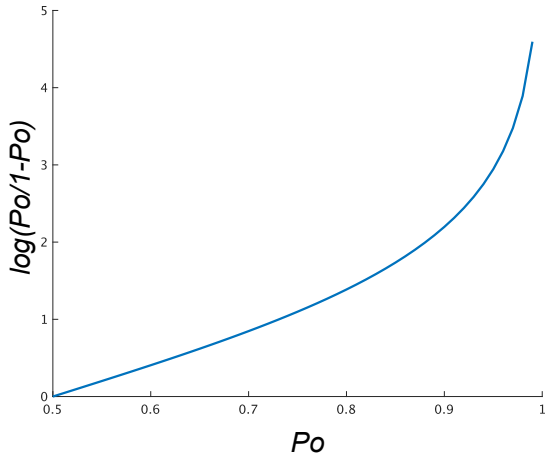


図 14: 報酬のロジット関数
報酬をP_oのロジット関数で表した。

5.2 報酬確率認識のための状態空間モデル

エージェントは報酬が原因wから確率的に生成されると認識しているとし、報酬確率λ_iは以下のように表した。

$$\lambda_i = f(w_i) \tag{10}$$

ここで、iは選択肢(本研究では、左右)を示し、$f(x) = 1/(1 + e^{-x})$とする。また、エージェントは、報酬確率が以下のようにランダムウォークで変動する曖昧な環境を想定しているとした。

$$w_{i,t} = w_{i,t-1} + \sigma_w \xi_{i,t} \tag{11}$$

ここで、t, $\xi_{i,t}$、σ_wはそれぞれ二者択一課題の試行、ガウスノイズ、ノイズ強度を表す。したがって、エージェントは、以下のような状態空間モデルで表される環境を想定して

いるとした。

$$P(\mathbf{w}_t|\mathbf{w}_{t-1}) = \mathcal{N}(\mathbf{w}_t|\mathbf{w}_{t-1}, \sigma_w^2\mathbf{I}) \tag{12}$$

$$P(o_t|\mathbf{w}_t, \mathbf{a}_t) = \prod_i \left[f(w_{i,t})^{o_t} \{1 - f(w_{i,t})\}^{1-o_t} \right]^{a_{i,t}} \tag{13}$$

ここで、$\mathbf{w}_t$とo_tは、それぞれステップtにおける両選択肢の報酬確率($\mathbf{w}_t = (w_{1,t}, w_{2,t})^T$)と報酬の有無の観測($o_t \in \{0,1\}$)を司る潜在変数を示している。また、$\mathbf{a}_t$はステップ$t$におけるエージェントの行動を表し、($\mathbf{a}_t \in \{(1,0)^T, (0,1)^T\}$)で表される。$\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \Sigma)$は平均$\boldsymbol{\mu}$、分散$\Sigma$のガウス分布を表し、$\sigma_w^2$は$\mathbf{w}$の遷移確率の分散を表す。$f(w_{i,t}) = 1/(1 + e^{-w_{i,t}})$はステップ$t$における選択肢$i$の報酬確率を表している。$\mathbf{w}_1$の初期分布は、$P(\mathbf{w}_1) = \mathcal{N}(\mathbf{w}_1|0, \kappa\mathbf{I})$で与えられ、ここで$\kappa$は分散を表す。

5.3 報酬確率認識のための自由エネルギー原理

エージェントの報酬確率の認識過程を、逐次ベイズ推定を用いて次のようにモデル化した。

$$P(\mathbf{w}_t|o_{1:t}, \mathbf{a}_{1:t}) \propto P(o_t|\mathbf{w}_t, \mathbf{a}_t) \int P(\mathbf{w}_t|\mathbf{w}_{t-1}) P(\mathbf{w}_{t-1}|\boldsymbol{o}_{1:t-1}, \mathbf{a}_{1:t-1}) d\mathbf{w}_{t-1} \tag{14}$$

$P(o_t|\mathbf{w}_t, \mathbf{a}_t)$が非ガウス型であるため, $\mathbf{w}_t$の事後分布$P(\mathbf{w}_t|o_{1:t}, \mathbf{a}_{1:t})$は非ガウス型となり、解析的に算出することができない。この問題を回避するために、ガウス分布で近似した単純な事後分布を導入し、以下のようにした。

$$Q(\mathbf{w}_t|\varphi_t) = \mathcal{N}(\mathbf{w}_t|\boldsymbol{\mu}_t, \boldsymbol{\Lambda}_t^{-1}) \fallingdotseq P(\mathbf{w}_t|o_{1:t}, \mathbf{a}_{1:t}) \tag{15}$$

ここで、$\varphi_t = \{\boldsymbol{\mu}_t, \boldsymbol{\Lambda}_t\}$であり、$\boldsymbol{\mu}_t$と$\boldsymbol{\Lambda}_t$はそれぞれ平均と精度を表す($\boldsymbol{\mu}_t = (\mu_{1,t}, \mu_{2,t})^T$; $\boldsymbol{\Lambda}_t = diag(p_{1,t}, p_{2,t})$)。また、$Q(\mathbf{w}_t|\varphi_t)$は、認識分布を表す。モデルエージェントは、$-lnP(o_t|o_{1:t-1})$で定義されるサプライズを最小化することによって、各時間ステップにおいて認識分布を更新する。ただし、サプライズは次のように分解できる。

$$-lnP(o_t|o_{1:t-1}) = \int Q(\mathbf{w}_t|\varphi_t) \ln \frac{Q(\mathbf{w}_t|\varphi_t)}{P(o_t, \mathbf{w}_t|o_{1:t-1}, \mathbf{a}_{1:t})} d\mathbf{w}_t - \mathrm{KL}[Q(\mathbf{w}_t|\varphi_t)||P(\mathbf{w}_t|o_{1:t}, \mathbf{a}_{1:t})] \tag{16}$$

ここで、$\mathrm{KL}[q(\mathbf{x})||p(\mathbf{x})]$は確率分布$q(\mathbf{x})$と$p(\mathbf{x})$の間の KL（Kullback-Leibler）距離を表す。KL 距離は非負であるため、第 1 項はサプライズの上限となる。

$$F(o_t, \varphi_t) = \int Q(\mathbf{w}_t|\varphi_t) \ln Q(\mathbf{w}_t|\varphi_t)\, d\mathbf{w}_t + \int Q(\mathbf{w}_t|\varphi_t) J(o_t, \mathbf{w}_t) d\mathbf{w}_t \tag{17}$$

は自由エネルギーと呼ばれる。ただし、$J(o_t, \mathbf{w}_t) = -lnP(o_t, \mathbf{w}_t|o_{1:t-1}, \mathbf{a}_{1:t})$である。自由エネルギーの第 1 項は、ガウス分布の負のエントロピーに対応する。

$$F_1 = \int Q(\mathbf{w}_t|\varphi_t) \ln Q(\mathbf{w}_t|\varphi_t)\, d\mathbf{w}_t \tag{18}$$

また、第 2 項は、次のように近似される。

$$F_2 = \int Q(\mathbf{w}_t|\varphi_t)J(o_t,\mathbf{w}_t)d\mathbf{w}_t$$
$$\cong \int Q(\mathbf{w}_t|\varphi_t)\left\{J(o_t,\boldsymbol{\mu}_t)+\frac{dJ}{dw}(\mathbf{w}_t-\boldsymbol{\mu}_t)+\frac{1}{2}\frac{d^2J}{dw^2}(\mathbf{w}_t-\boldsymbol{\mu}_t)^2\right\}d\mathbf{w}_t$$
$$= J(o_t,\mathbf{w}_t)|_{w_t=\mu_t}+\frac{1}{2}\frac{d^2J}{dw^2}\Big|_{w_t=\mu_t}\boldsymbol{\Lambda}_t^{-1} \quad (19)$$

ただし、$E(o_t,\mathbf{w}_t)$は$\boldsymbol{\mu}_t$の周りで 2 次のテイラー展開されている。各時間ステップにおいて、エージェントは$F(o_t,\varphi_t)$を最小化することで$\boldsymbol{\varphi}_t$を更新する。

5.4 自由エネルギーの計算

自由エネルギーは次のように導かれる。F_1は単純に次のようになる。

$$F_1 = \frac{1}{2}\ln 2\pi p_{1,t}^{-1}+\frac{1}{2}\ln 2\pi p_{2,t}^{-1}+const. \quad (20)$$

次にF_2を計算する。

$$P(o_t,\mathbf{w}_t|o_{1:t-1},\mathbf{a}_{1:t}) = P(o_t|\mathbf{w}_t,\mathbf{a}_t)\int P(\mathbf{w}_t|\mathbf{w}_{t-1})P(\mathbf{w}_{t-1}|o_{1:t-1},\mathbf{a}_{1:t-1})d\mathbf{w}_{t-1}$$

$$\cong P(o_t|\mathbf{w}_t,\mathbf{a}_t)\int P(\mathbf{w}_t|\mathbf{w}_{t-1})\,\mathrm{N}(\mathbf{w}_{t-1}|\boldsymbol{\mu}_{t-1},\boldsymbol{\Lambda}_{t-1}^{-1})d\mathbf{w}_{t-1} \quad (21)$$

この式の 2 行目では、近似した認識分布を前回の事後分布$P(\mathbf{w}_{t-1}|o_{1:t-1},\mathbf{a}_{1:t-1})$とし、次のように書くことができる。

$$P(o_t,\mathbf{w}_t|o_{1:t-1},\mathbf{a}_{1:t}) \cong$$
$$\prod_i\left[f(w_{i,t})^{o_t}\{1-f(w_{i,t})\}^{1-o_t}\right]^{a_{i,t}}N(w_{i,t}|\mu_{i,t-1},p_{i,t}^{-1}+\sigma_w^2) \quad (22)$$

$$E(o_t,\mathbf{w}_t)|_{w_t=\mu_t} = J_1(o_t,\mu_{1,t})+J_2(o_t,\mu_{2,t})+const \quad (23)$$

ここで、

$$J_i(o_t,\mu_{i,t}) = a_{i,t}\left[o_t\ln f(\mu_{i,t})+(1-o_t)\ln\{1-f(\mu_{i,t})\}\right]-\frac{1}{2}\frac{(\mu_{i,t}-\mu_{i,t-1})^2}{p_{i,t}^{-1}+\sigma_w^2}-\frac{1}{2}\ln(p_{i,t}^{-1}+\sigma_w^2), \quad (24)$$

とする。したがって、F_2はこの式を式(19)に代入することで算出される。まとめると、以下のようになる。

$$F(o_t,\varphi_t) = \sum_i\left\{J_i(o_t,\mu_{i,t})+\frac{1}{2}\frac{d^2J_i}{dw_{i,t}^2}\Big|_{w_{i,t}=\mu_{i,t}}p_{i,t}^{-1}+\frac{1}{2}\ln 2\pi p_{i,t}^{-1}\right\} \quad (25)$$

5.5 エージェントの認識の逐次更新

φ_tの更新則は、自由エネルギーを最小化することで導き出された。最適化された$p_{i,t}$は、$\partial F/\partial p_{i,t}^{-1} = 0$によって計算でき、次のようになる。

$$p_{i,t} = \left.\frac{d^2 J_i}{dw_{i,t}^2}\right|_{w_{i,t}=\mu_{i,t}} \tag{26}$$

$p_{i,t}$を式(25)に代入することで、$\mu_{i,t}$に関係なく、和の第 2 項は一定となる。したがって、$\mu_{i,t}$は、以下のように第 1 項のみを最小化することで更新される。

$$\mu_{i,t} = \mu_{i,t-1} - \alpha\delta_{i,a}\left.\frac{\partial J_i}{\partial \mu_{i,t}}\right|_{\mu_{i,t}=\mu_{i,t-1}} \tag{27}$$

ここで、α は学習率である。この 2 つの式から以下のような更新式が導かれる。

$$\mu_{i,t} = \mu_{i,t-1} + \alpha K_{i,t}\big(o_t - f(\mu_{i,t-1})\big) \tag{28}$$

$$p_{i,t} = K_{i,t}^{-1} + f(\mu_{i,t})(1 - f(\mu_{i,t})) \tag{29}$$

ここで、$K_{i,t} = (p_{i,t-1} + \sigma_w^{-2})/(p_{i,t-1}\sigma_w^{-2})$であり、カルマンゲインと呼ばれている。選択肢$i$が選択されなかった場合、両式の第 2 項は消滅し、その結果、認識$\mu_{i,t}$は変わらないが、その精度は低下する（すなわち、$p_{i,t+1} < p_{i,t}$）。一方で、選択された場合、認識は予測誤差によって更新され（すなわち、$o_t - f(\mu_{i,t})$）、その精度は向上する。ただし、認識された報酬確率に対する自信は、$w_{i,t}$空間ではなく、$\lambda_{i,t}$空間で評価する必要があり、

$\gamma_{i,t} = p_{i,t}/f'\big(\mu_{i,t}\big)^2$として表現できる。

5.6 期待純効用

期待純効用は次のように記述される。

$$U_t(\mathbf{a}_{t+1}) = c_t \cdot E_{P(o_{t+1}|\mathbf{a}_{t+1})}\big[\mathrm{KL}[Q(\mathbf{w}_{t+1}|o_{t+1},\mathbf{a}_{t+1})||Q(\mathbf{w}_{t+1}|\mathbf{a}_{t+1})]\big] + E_{P(o_{t+1}|\mathbf{a}_{t+1})}[R(o_{t+1})], \tag{30}$$

ここで、$R(o_{t+1}) = o_{t+1}\ln(P_o/(1-P_o))$である。式(30)の第一項と第二項はそれぞれ情報利得と期待報酬を表し、c_tは時間tにおける好奇心の強さを表す。好奇心のメタパラメータc_tを導入するというアイデアは強化学習分野[30]でも提案されている。以下では、期待純効用の導出方法を簡単に紹介する。現在の時刻tにおける自由エネルギーは次のように記述される。

$$F(o_t,\varphi_t) = E_{Q(\mathbf{w}_t|\varphi_t)}[\ln Q(\mathbf{w}_t|\varphi_t) - \ln P(o_t,\mathbf{w}_t|o_{1:t-1},\mathbf{a}_{1:t})] \tag{31}$$

これは、式(17)を書き換えたものとなる。ここで、行動$\mathbf{a}_{t+1}$を条件とした時刻 t+1 における未来の自由エネルギーを次のように表現する。

$$F(o_{t+1},\mathbf{a}_{t+1}) = E_{Q(\mathbf{w}_{t+1}|\varphi_t)}[\ln Q(\mathbf{w}_{t+1}|\varphi_t) - \ln P(o_{t+1},\mathbf{w}_{t+1}|o_{1:t},\mathbf{a}_{1:t+1})] \tag{32}$$

しかし、この式にo_{t+1}は将来の観測であるため、まだ観測されていない。この問題を解決するために、o_{t+1}にそれぞれ期待値をとって、次のようにする。

$$F(\mathbf{a}_{t+1}) = E_{Q(o_{t+1}|\mathbf{w}_{t+1},\mathbf{a}_{t+1})Q(\mathbf{w}_{t+1}|\varphi_t)}[\ln Q(\mathbf{w}_{t+1}|\varphi_t) - \ln P(o_{t+1},\mathbf{w}_{t+1}|o_{1:t},\mathbf{a}_{1:t+1})] \quad (33)$$

これは次のように書き換えることができる。

$$F(\mathbf{a}_{t+1}) = E_{Q(o_{t+1},\mathbf{w}_{t+1}|o_{1:t},\mathbf{a}_{t+1})}[\ln Q(\mathbf{w}_{t+1}|\varphi_t) - \ln P(\mathbf{w}_{t+1}|o_{1:t+1},\mathbf{a}_{1:t+1}) - \ln P(o_{t+1}|o_{1:t},\mathbf{a}_{1:t+1})] \quad (34)$$

また、$P(o_{t+1}|o_{1:t},\mathbf{a}_{1:t+1})$は、生成モデルによって、次のように計算することができる。

$$P(o_{t+1}|o_{1:t},\mathbf{a}_{1:t+1}) = \int P(o_{t+1}|\mathbf{w}_{t+1},\mathbf{a}_{1:t+1})P(\mathbf{w}_{t+1}|o_{1:t})d\mathbf{w}_{t+1}$$

$$\cong \int P(o_{t+1}|\mathbf{w}_{t+1},\mathbf{a}_{1:t+1})Q(\mathbf{w}_{t+1}|\varphi_t)d\mathbf{w}_{t+1} \quad (35)$$

ただし、$P(o_{t+1}|o_{1:t},\mathbf{a}_{1:t+1})$は、$P(o_{t+1})$と発見的に置き換えられたとし、$\ln P(o_{t+1})$を便宜的に報酬として扱う。フリストンらは、この式はさらに次のように変形した。

$$F(\mathbf{a}_{t+1}) = E_{Q(o_{t+1}|\mathbf{a}_{t+1})Q(\mathbf{w}_{t+1}|o_{1:t+1},\mathbf{a}_{t+1})}[\ln Q(\mathbf{w}_{t+1}|\varphi_t) - \ln Q(\mathbf{w}_{t+1}|o_{1:t+1},\mathbf{a}_{1:t+1}) - \ln P(o_{t+1})] \quad (36)$$

最後に、(36)式を変形し次のように期待自由エネルギー求めた。

$$F(\mathbf{a}_{t+1}) = E_{Q(o_{t+1}|\mathbf{a}_{t+1})}[-\mathrm{KL}[Q(\mathbf{w}_{t+1}|o_{t+1},\mathbf{a}_{t+1})||Q(\mathbf{w}_{t+1})] - \ln P(o_{t+1})] \quad (37)$$

ここで、第 1 項 KL ダイバージェンスはエージェントの認識が観測によって更新される程度を表すのに対し、第 2 項の$\ln P(o_{t+1})$はエージェントがo_{t+1}観測する事前選好として捉える事ができ、便宜的に報酬と解釈できる。

式（30）第 1 項の KL 距離における$\mathbf{w}_{t+1}$の事後分布と事前分布は、次のように導かれる。

$$Q(\mathbf{w}_{t+1}) = \int P(\mathbf{w}_{t+1}|\mathbf{w}_t)Q(\mathbf{w}_t)d\mathbf{w}_t \quad (38)$$

$$Q(\mathbf{w}_{t+1}|o_{t+1},\mathbf{a}_{t+1}) = \frac{P(o_{t+1}|\mathbf{w}_{t+1},\mathbf{a}_{t+1})Q(\mathbf{w}_{t+1}|\mathbf{a}_{t+1})}{P(o_{t+1}|\mathbf{a}_{t+1})} \quad (39)$$

この KL 距離はエージェントの認識が観測によって更新される度合いを表す。第 1 項は次のように計算することが出来る。

$$P(o_{t+1}|\mathbf{a}_{t+1}) = \int P(o_{t+1}|\mathbf{w}_{t+1},\mathbf{a}_{t+1})Q(\mathbf{w}_{t+1}|\mathbf{a}_{t+1})d\mathbf{w}_{t+1} \quad (40)$$

第 2 項では、報酬をo_{t+1}の期待する確率として定量的に解釈しており、以下のように表現した。

$$P(o_{t+1}) = P_o^{o_{t+1}}(1-P_o)^{(1-o_{t+1})} \quad (41)$$

ここで、P_oは報酬の存在する望ましい確率を示す。報酬の確率論的解釈 [10,28] によれば、報酬の有無はそれぞれ$\ln P_o$と$\ln(1-P_o)$で評価することができる。

本研究で提案した期待効用はフリストンが提案した期待自由エネルギーと以下の点で異なる。すなわち、意思決定を期待自由エネルギーの最大化としてモデル化したため、期待自由エネルギーの符号変更し、期待純効用と表現した。また、非合理的な意思決定を表現するために、好奇心メタパラメータc_tを導入したほか、報酬を$\ln\{P_o/(1-P_o)\}$と定式化した。

5.7 行動選択のモデル化

エージェントは、以下のように確率的に期待純効用の高い選択肢を選択する。

$$P(\mathbf{a}_{t+1}) = \frac{exp\big(\beta U(\mathbf{a}_{t+1})\big)}{\sum_{\mathbf{a}} exp\big(\beta U(\mathbf{a})\big)} \tag{42}$$

ここで、$U(\mathbf{a}_{t+1})$は行動$\mathbf{a}_{t+1}$の期待純効用を示し、式（42）は式（2）と同じである。式(42)を導くために、確率的行動に対する期待純効用の期待値を以下のように記述する。

$$E[U] = E_{Q(\mathbf{a}_{t+1})}[U(\mathbf{a}_{t+1}) - \beta^{-1}\ln Q(\mathbf{a}_{t+1})] \tag{43}$$

ここで β は逆温度を示し、第 2 項には作用確率のエントロピー的制約が導入されている。この式は次のように書き換えることができる。

$$E[U] = -\beta^{-1}KL[Q(\mathbf{a}_{t+1})\|\exp(\beta U(\mathbf{a}_{t+1}))/Z] + \beta^{-1}\ln Z \tag{44}$$

ここで、Z は正規化定数を示す。したがって、$Q(\mathbf{a}_{t+1})$に対するそれぞれの最大化は、式(42)に示すように最適な行動確率を導く。

5.8 期待純効用の別の記述

比較のため、第 2 項に時間依存のメタパラメータを導入して、代替的な期待純効用を検討すると、次のようになる。

$$\begin{aligned} U(\mathbf{a}_{t+1}) = E_{P(o_{t+1}|\mathbf{a}_{t+1})}\big[\mathrm{KL}[Q(\mathbf{w}_{t+1}|o_{t+1},\mathbf{a}_{t+1})||Q(\mathbf{w}_{t+1}|\mathbf{a}_{t+1})]\big] \\ +d_t \cdot E_{P(o_{t+1}|\mathbf{a}_{t+1})}[R(o_{t+1})] \end{aligned} \tag{45}$$

この場合、d_tが高いエージェントはより搾取的な行動をとり、$d_t = 0$のエージェントは期待される情報利得によってより探索的な行動をとることになる。

5.9 期待純効用の導出

ここでは、期待純効用の計算を紹介する。式(30)の第 1 項の KL 距離は次のように変形できる。

$$E_{p(o_{t+1}|\mathbf{a}_{t+1})}\big[KL[Q(\mathbf{w}_{t+1}|o_{t+1},\mathbf{a}_{t+1})||Q(\mathbf{w}_{t+1}|\mathbf{a}_{t+1})]\big] = \mathrm{H}(o_{t+1}) - \mathrm{H}(o_{t+1}|\mathbf{w}_{t+1}) \tag{46}$$

$$\mathrm{H}(o_{t+1}|\mathbf{w}_{t+1}) = \mathrm{E}_{P(o_{t+1}|\mathbf{w}_{t+1},\mathbf{a}_{t+1})Q(\boldsymbol{w}_{t+1}|\boldsymbol{a}_{t+1})}[-\ln P(o_{t+1}|\mathbf{w}_{t+1},\mathbf{a}_{t+1})] \tag{47}$$

$$\mathrm{H}(o_{t+1}) = \mathrm{E}_{P(o_{t+1}|\mathbf{a}_{t+1})}[-\ln P(o_{t+1}|\mathbf{a}_{t+1})] \tag{48}$$

条件付きエントロピーH$(o_{t+1}|\mathbf{w}_{t+1})$は、式（13）を式（47）に代入して、次のように算出することができる。

$$\mathrm{H}(o_{t+1}|\mathbf{w}_{t+1}) = -E_{Q(\mathbf{w}_{t+1}|\mathbf{a}_{t+1})}\left[\sum\nolimits_i a_{i,t+1} g\big(w_{i,t+1}\big)\right] \tag{49}$$

ただし、

$$g(w) = f(w)\ln f(w) + \big(1-f(w)\big)\ln\big(1-f(w)\big) \tag{50}$$

である。ここで、この式を 2 次のテイラー展開を用いて近似的に計算すると、次のようになる。

$$\mathrm{H}(o_{t+1}|\mathbf{w}_{t+1}) \cong -E_{Q(\mathbf{w}_{t+1}|\mathbf{a}_{t+1})}\left[\sum\nolimits_i a_{i,t+1}\left\{\begin{array}{c} g(\mu_{i,t+1}) + \dfrac{\partial g}{\partial w_{i,t+1}}(w_{i,t+1}-\mu_{i,t+1}) \\ +\dfrac{1}{2}\dfrac{\partial^2 g}{\partial w_{i,t+1}^2}(w_{i,t+1}-\mu_{i,t+1})^2 \end{array}\right\}\right] \quad (51)$$

これは以下のように記述できる。

$$\mathrm{H}(o_{t+1}|\mathbf{w}_{t+1}) = -\sum\nolimits_i a_{i,t+1}\left[\begin{array}{c} f(\mu_{i,t+1})\ln f(\mu_{i,t+1}) + \left(1-f(\mu_{i,t+1})\right)\ln\left(1-f(\mu_{i,t+1})\right) \\ +\dfrac{1}{2}\left\{f(\mu_{i,t+1})\left(1-f(\mu_{i,t+1})\right)\left(1+\left(1-2f(\mu_{i,t+1})\right)\ln\dfrac{f(\mu_{i,t+1})}{1-f(\mu_{i,t+1})}\right)\right\} \\ (p_{i,t}^{-1}+p_w^{-1}) \end{array}\right] \quad (52)$$

また、$\mathrm{H}(o_{t+1})$は以下のとおりである。

$$\mathrm{H}(o_{t+1}) = -\sum\nolimits_i a_{i,t+1}\{P(o_{t+1}=0|\mathbf{a}_{t+1})\ln P(o_{t+1}=0|\mathbf{a}_{t+1}) + P(o_{t+1}=1|\mathbf{a}_{t+1})\ln P(o_{t+1}=1|\mathbf{a}_{t+1})\} \quad (53)$$

ただし、

$$\begin{aligned} P(o_{t+1}|\mathbf{a}_{t+1}) &= \int P(o_{t+1}|\mathbf{w}_{t+1},\mathbf{a}_{t+1})Q(\mathbf{w}_{t+1}|\mathbf{a}_{t+1})d\mathbf{w}_{t+1} \\ &= \int \prod\nolimits_i \left\{f(w_{i,t+1})^{o_{t+1}}\left(1-f(w_{i,t+1})\right)^{1-o_{t+1}}\right\}^{a_{i,t+1}} Q(\mathbf{w}_{t+1}|\mathbf{a}_{t+1})d\mathbf{w}_{t+1} \\ &= \prod\nolimits_i \left[\begin{array}{c} f(\mu_{i,t+1})^{o_{t+1}}\left(1-f(\mu_{i,t+1})\right)^{1-o_{t+1}} \\ +1^{o_{t+1}}(-1)^{1-o_{t+1}}\dfrac{1}{2}f(\mu_{i,t+1})\{1-f(\mu_{i,t+1})\}\{1-2f(\mu_{i,t+1})\}(p_{i,t}^{-1}+p_w^{-1}) \end{array}\right]^{a_i,t+1} \end{aligned} \quad (54)$$

また、期待純効用の第 2 項（式(36)）は、次のように計算される。

$$\begin{aligned} E_{P(o_{t+1}|\mathbf{a}_{t+1})}[\ln P(o_{t+1})] &= E_{P(o_{t+1}|\mathbf{a}_{t+1})}[o_{t+1}\ln P_o + (1-o_{t+1})\ln(1-P_o)] \\ &= P(o_{t+1}=0|\mathbf{a}_{t+1})\ln(1-P_o) + P(o_{t+1}=1|\mathbf{a}_{t+1})\ln(1-P_o) \end{aligned} \quad (55)$$

5.10 観測者視点での状態空間モデル(観測者−SSM)

本研究では観測者の視点から、エージェントの潜在的な内部状態の時間的遷移と行動の発生を記述する観測者-SSM を構築した（図 5)。エージェントは内部状態、すなわち好奇心の強さ、認識された報酬確率、およびその自信に基づいて行動すると仮定した。また、好奇心の強さは、ランダムウォークとして時間的に変化すると仮定した。

$$c_t = c_{t-1} + \epsilon\zeta_t \quad (56)$$

ここで、ζ_tは平均がゼロで分散が 1 のホワイトノイズを示し、ϵはそのノイズ強度を示している。その他の内部状態、すなわちμ_iとp_iは、式（28）と式（29）のように更新されるとした。内部状態の遷移は、以下の確率分布で表される。

$$P(\mathbf{z}_t|\mathbf{z}_{t-1}) = \mathcal{N}(\mathbf{z}_t|\mathbf{F}(\mathbf{z}_{t-1},\mathbf{a}_{t-1}),\boldsymbol{\Gamma}) \quad (57)$$

$$\mathbf{F}(\mathbf{z}_{t-1}, \mathbf{a}_{t-1}) = \begin{bmatrix} 0 \\ h(\mu_{1,t-1}, p_{1,t-1}, o_t, a_1) \\ h(\mu_{2,t-1}, p_{2,t-1}, o_t, a_2) \\ k(\mu_{1,t-1}, p_{1,t-1}, a_1) \\ k(\mu_{2,t-1}, p_{2,t-1}, a_2) \end{bmatrix} \tag{58}$$

ここで、$\mathbf{z}_t = (c_t, \boldsymbol{\mu}_t^T, \boldsymbol{p}_t^T)^{\mathrm{T}};\ \boldsymbol{\Gamma} = \epsilon^2 diag(1,0,0,0,0)$とした。また、$h(\mu_{i,t-1}, p_{i,t-1}, o_t, a_i)$と$k(\mu_{i,t-1}, p_{i,t-1}, a_i)$ はそれぞれ式(28)と式(29)の右辺を表す。$\boldsymbol{\Gamma}$と$diag(\mathbf{x})$は、それぞれ分散共分散行列と対角成分を$\mathbf{x}$とする正方行列を表す。また、エージェントは、期待純効用に基づいて行動$\mathbf{a}_{t+1}$を以下のように選択するものとした。

$$P(\mathbf{a}_{t+1}) = \frac{\exp\big(\beta U(\mathbf{a}_{t+1})\big)}{\sum_{\mathbf{a}} \exp\big(\beta U(\mathbf{a})\big)} \tag{59}$$

なお、報酬は以下の確率分布で得られた。

$$P(o_t|\mathbf{a}_t) = \prod\nolimits_i \left\{\lambda_{i,t}^{o_t}\big(1-\lambda_{i,t}\big)^{1-o_t}\right\}^{a_{i,t}} \tag{60}$$

5.11 二者択一課題へのQ学習モデルのあてはめ

二者択一課題における選択もQ学習モデルによってモデル化された。i番目の選択肢の報酬予測$Q_{i,t}$は以下のように更新される。

$$Q_{i,t} = Q_{i,t-1} + \alpha_{t-1}\big(r_t a_{i,t-1} - Q_{i,t-1}\big) \tag{61}$$

ここで、α_tは試行tにおける学習率を示す。得られた選択肢の報酬予測に基づき、エージェントは以下のソフトマックス関数に従って行動を選択する。

$$P\big(a_{i,t}=1\big) = \frac{exp\big(B_t Q_{i,t}\big)}{\sum_i exp\big(B_t Q_{i,t}\big)} \tag{62}$$

ここで、B_tは行動選択のランダム性を制御する試行tにおける逆温度を示す。Q学習における時間依存のパラメータα_tとB_tは、行動データ[38]から推定した。これらのパラメータは時間的にランダムウォークすると仮定した。

$$\theta_t = \theta_{t-1} + \epsilon_\theta \zeta_{\theta,t} \tag{63}$$

ここで、$\theta \in \{\alpha, B\}$、$\zeta_{\theta,t}$は平均がゼロ、分散が1のホワイトノイズを示し、ϵ_θはそのノイズ強度を示している。すなわち、内部状態の遷移は、以下の確率分布で表される。

$$P(\mathbf{z}_t|\mathbf{z}_{t-1}) = \mathcal{N}(\mathbf{z}_t|\mathbf{F}(\mathbf{z}_{t-1}, \mathbf{a}_{t-1}), \boldsymbol{\Gamma}) \tag{64}$$

$$\mathbf{F}(\mathbf{z}_{t-1}, \mathbf{a}_{t-1}) = \begin{bmatrix} 0 \\ 0 \\ h_1(\alpha_t, Q_{1,t-1}, \mathbf{a}_{t-1}) \\ h_2(\alpha_t, Q_{2,t-1}, \mathbf{a}_{t-1}) \end{bmatrix} \tag{65}$$

ここで、$\mathbf{z}_t = \big(\alpha_t, B_t, Q_{1,t}, Q_{2,t}\big)^{\mathrm{T}};\ \boldsymbol{\Gamma} = \epsilon^2 diag(1,1,0,0);\ h_i(\alpha_t, Q_{i,t-1}, \mathbf{a}_t)$ は式(61)の右辺を表し、$\boldsymbol{\Gamma}$と$diag(\mathbf{x})$はそれぞれ分散共分散行列と対角成分を$\mathbf{x}$とする正方行列を表す。

5.12 粒子フィルタとカルマンバックワードアルゴリズムによる iFEP の実装

観測者-SSM に基づき、1 から $T(x_{1:T})$までの全ての観測値を与えられたエージェントz_tの潜在的な内部状態の事後分布、すなわち$P(z_t|x_{1:T})$をベイズ推定した。ベイズ推定において、それぞれフィルタリング、スムージングと呼ばれるフォワード、バックワードアルゴリズムを用いて推定された。フィルタリングでは、tまでの観測値$(x_{1:t})$が与えられたz_tの事後分布を以下のように、順方向に逐次更新していく。

$$P(\mathbf{z}_t|\mathbf{x}_{1:t}) \propto P(\mathbf{x}_t|\mathbf{z}_t,\theta)\int P(\mathbf{z}_t|\mathbf{z}_{t-1},\theta)P(\mathbf{z}_{t-1}|\mathbf{x}_{1:t-1})d\mathbf{z}_{t-1} \tag{66}$$

ここで、$\mathbf{x}_t = (\mathbf{a}_t^T, o_t)^{\mathrm{T}}$,$\theta = \{\sigma^2, \alpha, P_o\}$とする。$\mathbf{z}_1$の事前分布は以下のように設定した。

$$P(\mathbf{z}_1) = \left[\prod\nolimits_i \mathcal{N}\left(\mu_{i,1}\middle|\mu_0,\sigma_\mu^2\right)Gam\left(p_{i,1}\middle|a_g,b_g\right)\right]Uni(c_1|a_u,b_u) \tag{67}$$

ただし、μ_0とσ_μ^2は平均と分散を示し、$Gam(x|a_g,b_g)$は形状パラメータa_gとスケールパラメータb_gのガンマ分布、$Uni(x|a_u,b_u)$はa_uからb_uの間の一様分布である。$P(\mathbf{z}_t|\mathbf{x}_{1:t})$は非線形であることから解析的に導出できないため粒子フィルター[43] を用いて逐次算出した。粒子フィルターで$P(\mathbf{z}_t|\mathbf{x}_{1:t})$を算出した後、すべての観測値$(\mathbf{x}_{1:T})$ が与えられた$\mathbf{z}_t$の事後分布$P(\mathbf{z}_t|\mathbf{x}_{1:T})$は、次のように逆方向に逐次更新した。

$$\begin{aligned} P(\mathbf{z}_t|\mathbf{x}_{1:T}) &= \int P(\mathbf{z}_{t+1}|\mathbf{x}_{1:T})P(\mathbf{z}_t|\mathbf{z}_{t+1},\mathbf{x}_{1:t},\theta)d\mathbf{z}_{t+1} \\ &= \int P(\mathbf{z}_{t+1}|\mathbf{x}_{1:T})\frac{P(\mathbf{z}_{t+1}|\mathbf{z}_t,\theta)P(\mathbf{z}_t|\mathbf{x}_{1:t},\theta)}{\int P(\mathbf{z}_{t+1}|\mathbf{z}_t,\theta)P(\mathbf{z}_t|\mathbf{x}_{1:t},\theta)d\mathbf{z}_t}d\mathbf{z}_{t+1} \end{aligned} \tag{68}$$

しかし、粒子フィルタの粒子アンサンブルで表現した$P(\mathbf{z}_T|\mathbf{x}_{1:T})$が非ガウスであることに加え、$P(\mathbf{z}_{t+1}|\mathbf{z}_t,\theta)$における$\mathbf{z}_t$と$\mathbf{z}_{t+1}$が非線形な関係（式（57））のため、式(63)を計算することは困難である。したがって、$P(\mathbf{z}_t|\mathbf{x}_{1:t})$を$\mathcal{N}(\mathbf{z}_t|\mathbf{m}_t,\mathbf{V}_t)$として近似し、$P(\mathbf{z}_t|\mathbf{z}_{t-1},\theta)$は以下の通り線形化した。ただし、$\mathbf{m}_t$と$\mathbf{V}_t$は$t$における粒子のサンプル平均とサンプル分散を表す。

$$P(\mathbf{z}_t|\mathbf{z}_{t-1},\theta) \cong \mathcal{N}(\mathbf{z}_t|\mathbf{A}\mathbf{z}_{t-1}+\mathbf{b},\boldsymbol{\Gamma}), \tag{69}$$

$$\mathbf{A} = \left.\frac{\partial \mathbf{F}(\mathbf{z}_{t-1},\mathbf{a}_{t-1})}{\partial \mathbf{z}_{t-1}}\right|_{\mathbf{m}_t}, \tag{70}$$

$$\mathbf{b} = F(\mathbf{m}_t,\mathbf{a}_{t-1}) - \mathbf{A}\mathbf{m}_t, \tag{71}$$

ここで、$\mathbf{A}$はヤコビアン行列を表す。これらの近似により、式(68)の積分が計算可能になり、事後分布$P(\mathbf{z}_t|\mathbf{x}_{1:T})$は、ガウス分布により次のように計算することができる。

$$P(\mathbf{z}_t|\mathbf{x}_{1:T}) = \mathcal{N}\left(\mathbf{z}_t|\widehat{\mathbf{m}}_t,\widehat{\mathbf{V}}_t\right). \tag{72}$$

この平均と分散は、カルマンバックワードアルゴリズム [44] によって解析的に更新された。

$$\widehat{\mathbf{m}}_t = \mathbf{m}_t + \mathbf{J}_t\{\widehat{\mathbf{m}}_{t+1} - (\mathbf{A}\mathbf{m}_t+\mathbf{b})\}, \tag{73}$$

$$\widehat{\mathbf{V}}_t = \mathbf{V}_t + \mathbf{J}_t\{\widehat{\mathbf{m}}_{t+1} - (\mathbf{A}\mathbf{V}_t\mathbf{A}^{\mathrm{T}} + \boldsymbol{\Gamma})\}\mathbf{J}_t^{\mathrm{T}}, \tag{74}$$

ただし、$\mathbf{J}_t$は以下のとおり。

$$\mathbf{J}_t = \mathbf{V}_t\mathbf{A}^{\mathrm{T}}\left(\mathbf{A}\mathbf{V}_t\mathbf{A}^{\mathrm{T}} + \boldsymbol{\Gamma}\right)^{-\mathbf{1}}. \tag{75}$$

5.13 モデル識別の不可能性

ReCU と Q 学習モデルにおいて、行動の選択は、以下のように同じソフトマックス関数で定式化されている。

$$P\left(a_{i,t} = 1\right) = \frac{\exp\left(\beta\left(E[R_{i,t}] + c_t E[Info_{i,t}]\right)\right)}{\sum_j \exp\left(\beta\left(E[R_{j,t}] + c_t E[Info_{j,t}]\right)\right)}. \tag{76}$$

$$P\left(a_{i,t} = 1\right) = \frac{\exp\left(\beta_t Q_{i,t}\right)}{\sum_i \exp\left(\beta_t Q_{i,t}\right)}, \tag{77}$$

式(76)と式(77)は、それぞれ ReCU と Q 学習のモデルに対応する。これらの式には、時間依存のメタパラメータであるc_tとβ_tが含まれている。両モデルとも、時間依存のメタパラメータを調整することで、実際の行動データに対する適合度（＝尤度）を自在にに改善することができることから、ReCU モデルと Q 学習モデルの識別は本質的に不可能である。

5.14 iFEP におけるパラメータの推定

ReCU モデルは、σ_w^2、α、β、P_o、ϵという複数のパラメータを含んでいる。推定では、ϵの値を人工データでの推定に最適であった 1 に設定した（図 7）。また、期待純効用ではβと$\ln P_o/(1-P_o)$を乗じることから、P_oとβの両方を推定することは不可能であり、$\ln P_o/(1-P_o) = 1$と仮定した（式（30，42 参照））。また、βc_tは、期待純効用にβがかけられるため、$\hat{c}_t = \beta c_t$とし、$\hat{c}_t$を潜在変数として扱った（式（30, 42）参照）。したがって、c_tの推定は、推定$\hat{c}_t$を推定βで割ることで得られる。こうしたもとで、推定すべきハイパーパラメータはσ_w^2、α、βである分かる。これらのパラメータ$\theta = \{\sigma_w^2, \alpha, \beta\}$を推定するために、観測者-SSM を自己組織化状態空間モデル [45] に拡張し、θ を一定の潜在変数として扱えるようにした。

$$P(\mathbf{z}_t, \theta|\mathbf{x}_{1:t}) \propto P(\mathbf{x}_t|\mathbf{z}_t)\int P(\mathbf{z}_t|\mathbf{z}_{t-1}, \theta)P(\mathbf{z}_{t-1}, \theta|\mathbf{x}_{1:t-1})d\mathbf{z}_{t-1} \tag{78}$$

ここで、$P(\theta) = Uni(\sigma^2|a_\sigma, b_\sigma)Uni(\alpha|a_\alpha, b_\alpha)\mathcal{N}(\beta|m_\beta, v_\beta)$である。粒子フィルタを用いて事後$P(\mathbf{z}_t, \theta|\mathbf{x}_{1:t})$を順次計算するために、10 万個の粒子を用い、ランダムにサンプリングした初期値から更新しないパラメータ θ を追加して全粒子の状態ベクトルを推定した。ただし、この推定に用いたハイパーパラメータ値は、$\mu_0 = 0$、$\sigma_\mu^2 = 0.01^2$、$a_g = 10$、$b_g = 0.001$、$a_u = -15$、$b_u = 15$、$a_\sigma = 0.2$、$b_\sigma = 0.7$、$a_\alpha = 0.04$、$b_\alpha = 0.06$、$a_\beta = 0$、$b_\beta = 50$で、人工データで正しく推定できたパラメータとして発見的に与えた（図 6）。

5.15 モンテカルロ・シミュレーションによる統計的検定

図 11 では、図 8 で推定した負の好奇心について統計的に検証した。帰無仮説は、「好奇心を持たないエージェント（すなわち$c_t = 0$）は、報酬確率の認識によってのみ選択を決定する」というものである。帰無仮説のもと、図 8 と同じ実験条件でモデルシミュレーションを 1,000 回繰り返し、それぞれについて iFEP を用いて好奇心を推定した。推定された好奇心の時間平均を検定統計量として採用し、検定統計量の帰無分布をプロットした。ラット行動の推定好奇心と比較し、片側左側検定の p 値を算出した。

6.データおよびコードの公開

6.1 本研究で用いたデータ

図表を作成するために用いたデータは以下のウェブサイトで公開している。また本研究で用いたラットの行動データは公開データを使用しており(伊藤、銅谷 2009)[37]、以下の URL から取得可能である。

図 2,4,8,10	https://www.nature.com/articles/s43588-023-00439-w#Sec32
図 6,7,9,11 12,13,14	https://www.nature.com/articles/s43588-023-00439-w#additional-information
行動データ	https://groups.oist.jp/ja/ncu/data

6.2 本研究で用いたコード

本研究でのシミュレーションおよびデータ解析において、MATLAB(R2020b)を用いた。また、用いたコードは GitHub からダウンロード可能である。また、本論文作成時点でのコードは Zenodo で公開している [46]。

GitHub	https://github.com/YukiKonaka/Konaka_Honda_2023
Zenodo	https://zenodo.org/record/7722905

7.引用文献

1. Helmholtz, H. Handbuch der Physiologischen Optik. 1867 (1867).
2. Yuille, A. & Kersten, D. Vision as Bayesian inference: analysis by synthesis? *Trends Cogn. Sci.* **10**, 301–308 (2006).
3. Friston, K. A theory of cortical responses. *Philos. Trans. R. Soc. B Biol. Sci.* **360**, 815–836 (2005).
4. Friston, K., Kilner, J. & Harrison, L. A free energy principle for the brain. *J. Physiol. Paris* **100**, 70–87 (2006).
5. Friston, K. The free-energy principle: A unified brain theory? *Nat. Rev. Neurosci.* **11**, 127–138 (2010).
6. Statistics, M. On a Measure of the Information Provided by an Experiment Author (s): D . V . Lindley Source : The Annals of Mathematical Statistics , Vol . 27 , No . 4 (Dec ., 1956), pp . 986-1005 Published by : Institute of Mathematical Statistics Stable URL : http. *Statistics (Ber).* **27**, 986–1005 (2009).
7. MacKay, D. J. C. Information-Based Objective Functions for Active Data Selection. *Neural Comput.* **4**, 590–604 (1992).
8. Berger, J. O. Statistical Decision Theory and Bayesian Analysis (Springer Series in Statistics). *Springer Series in Statistics* (2011).
9. Friston, K. *et al.* Active inference and epistemic value. *Cogn. Neurosci.* **6**, 187–214 (2015).
10. Friston, K., FitzGerald, T., Rigoli, F., Schwartenbeck, P. & Pezzulo, G. Active Inference: A Process Theory. *Neural Comput.* **29**, 1–49 (2017).
11. Attias, H. Planning by probabilistic inference. *Proc. 9th Int. Work. Artif. Intell. Stat.* (2003).
12. Botvinick, M. & Toussaint, M. Planning as inference. *Trends Cogn. Sci.* **16**, 485–488 (2012).
13. Kaplan, R. & Friston, K. J. Planning and navigation as active inference. *Biol. Cybern.* **112**, 323–343 (2018).
14. Matsumoto, T. & Tani, J. Goal-directed planning for habituated agents by active inference using a variational recurrent neural network. *Entropy* **22**, (2020).
15. Millett, J. D. & Simon, H. A. Administrative Behavior: A Study of Decision-Making Processes in Administrative Organization. *Polit. Sci. Q.* **62**, 621 (1947).
16. Dubey, R. & Griffiths, T. L. Understanding exploration in humans and machines by formalizing the function of curiosity. *Curr. Opin. Behav. Sci.* **35**, 118–124 (2020).
17. Kidd, C. & Hayden, B. Y. The Psychology and Neuroscience of Curiosity. *Neuron* **88**,

449–460 (2015).

18. Klein, U. & Nowak, A. J. Characteristics of patients with autistic disorder (AD) presenting for dental treatment: a survey and chart review. *Spec. care Dent. Off. Publ. Am. Assoc. Hosp. Dent. Acad. Dent. Handicap. Am. Soc. Geriatr. Dent.* **19**, 200–207 (1999).
19. Lockner, D. W., Crowe, T. K. & Skipper, B. J. Dietary Intake and Parents' Perception of Mealtime Behaviors in Preschool-Age Children with Autism Spectrum Disorder and in Typically Developing Children. *J. Am. Diet. Assoc.* **108**, 1360–1363 (2008).
20. Schreck, K. A. & Williams, K. Food preferences and factors influencing food selectivity for children with autism spectrum disorders. *Res. Dev. Disabil.* **27**, 353–363 (2006).
21. Esposito, M. *et al.* Sensory Processing, Gastrointestinal Symptoms and Parental Feeding Practices in The Explanation of Food Selectivity: Clustering Children with and Without Autism. *Int. J. Autism Relat. Disabil.* **2**, (2019).
22. Hobson, R. P. Autism and the Development of Mind. *Essays Dev. Psychol.* (1993).
23. Burke, R. Personalized recommendation of PoIs to people with autism. *Commun. ACM* **65**, 100 (2022).
24. Ghanizadeh, A. Educating and counseling of parents of children with attention-deficit hyperactivity disorder. *Patient Educ. Couns.* **68**, 23–28 (2007).
25. Sedgwick, J. A., Merwood, A. & Asherson, P. The positive aspects of attention deficit hyperactivity disorder: a qualitative investigation of successful adults with ADHD. *ADHD Atten. Deficit Hyperact. Disord.* **11**, 241–253 (2019).
26. Sutton, R. S. & Barto, A. G. *Reinforcement Learning: An Introduction*. (MIT Press, 1998).
27. Schwartenbeck, P. *et al.* Computational mechanisms of curiosity and goal-directed exploration. *Elife* **8**, 1–45 (2019).
28. Friston, K. *et al.* Active inference and epistemic value. *Cogn. Neurosci.* **6**, 187–214 (2015).
29. Millidge, B., Tschantz, A. & Buckley, C. L. Whence the expected free energy? *Neural Comput.* **33**, 447–482 (2021).
30. Houthooft, R. *et al.* VIME: Variational information maximizing exploration. *Adv. Neural Inf. Process. Syst.* **0**, 1117–1125 (2016).
31. Redshaw, R. & McCormack, L. "Being ADHD": a Qualitative Study. *Adv. Neurodev. Disord.* **6**, 20–28 (2022).
32. Smith, R. *et al.* Greater decision uncertainty characterizes a transdiagnostic patient sample during approach-avoidance conflict: A computational modelling approach. *J. Psychiatry Neurosci.* **46**, E74–E87 (2021).

33. Smith, R. *et al.* Long-term stability of computational parameters during approach-avoidance conflict in a transdiagnostic psychiatric patient sample. *Sci. Rep.* **11**, 1–13 (2021).
34. Schwartenbeck, P. & Friston, K. Computational phenotyping in psychiatry: A worked example. *eNeuro* **3**, 47 (2016).
35. Daunizeau, J. *et al.* Observing the observer (I): Meta-bayesian models of learning and decision-making. *PLoS One* **5**, (2010).
36. Patzelt, E. H., Hartley, C. A. & Gershman, S. J. Computational Phenotyping: Using Models to Understand Individual Differences in Personality, Development, and Mental Illness. *Personal. Neurosci.* **1**, (2018).
37. Ito, M. & Doya, K. Validation of decision-making models and analysis of decision variables in the rat basal ganglia. *J. Neurosci.* **29**, 9861–9874 (2009).
38. Samejima, K., Doya, K., Ueda, Y. & Kimura, M. Estimating internal variables and parameters of a learning agent by a particle filter. *Adv. Neural Inf. Process. Syst.* (2004).
39. Samejima, K., Ueda, Y., Doya, K. & Kimura, M. Neuroscience: Representation of action-specific reward values in the striatum. *Science (80-.).* **310**, 1337–1340 (2005).
40. Mizoguchi, H. *et al.* Insular neural system controls decision-making in healthy and methamphetamine-treated rats. *Proc. Natl. Acad. Sci. U. S. A.* **112**, E3930–E3939 (2015).
41. Katahira, K. The statistical structures of reinforcement learning with asymmetric value updates. *J. Math. Psychol.* **87**, 31–45 (2018).
42. Ortega, P. A. & Braun, D. A. Thermodynamics as a theory of decision-making with information-processing costs Subject Areas : Author for correspondence : *Proc. R. Soc. London. Part A* **469**, 20120683 (2013).
43. Kitagawa, G. A Monte Carlo Filtering and Smoothing Method for Non_Gaussian Nonlinear State Space Models. in (Proceedings of the 2nd U,S,-Japan Joint Seminar on Statisitical Time Series Analysis, 1993).
44. Bishop, C. M. *Pattern recognition and machine learning*. (New York : Springer, [2006] ©2006).
45. Journal, S., Statistical, A. & Sep, N. A Self-Organizing State-Space Model Author (s): Genshiro Kitagawa Published by : Taylor & Francis , Ltd . on behalf of the American Statistical Association Stable URL : http://www.jstor.org/stable/2669862 All use subject to http://about.jstor.org/terms. *Am. Stat.* **53**, 326–331 (1999).
46. Konaka, Y. & Naoki, H. Codes for Konaka and Honda 2023. (2023) doi:10.5281/zenodo.7722905.

8.謝辞

本論文の執筆にあたりお世話になった皆様に心から感謝申し上げます。まず、指導教員である本田先生におかれては、修士課程から約 5 年間、ご指導賜りましたこと感謝いたします。とくに、社会人学生として本学に入学して以降は、時間制約が厳しいもとで、夜遅い時間であっても丁寧にご指導頂きありがとうございました。また、本田研究室のスタッフ・学生の皆様にも心から感謝いたします。最後に、学位取得にむけて様々な困難があるなか、温かく励ましてくれた家族、親族に深く感謝いたします。

以上、多くの方々のご協力と支援のおかげで、私はこの博士論文を完成することができました。関わったすべての方々に改めて心から感謝申し上げます。