# 多様な環境の学習におけるACT-Rを用いた内発的動機づけのモデル
# Model with Intrinsic Motivation Using ACT-R in Learning Diverse Environments

長島 一真[†]，森田 純哉[‡]，竹内 勇剛[‡]
Kazuma Nagashima, Junya Morita, Yugo Takeuchi
[†] 静岡大学
Shizuoka University
nagashima.kazuma.16@shizuoka.ac.jp, j-morita@inf.shizuoka.ac.jp, takeuchi@inf.shizuoka.ac.jp

## 概要

人工エージェントが多様な環境を学習するためには，内発的動機に基づく報酬が必要である．これまでに，エージェントの内発的動機づけの研究が行われてきたが，統合的なアーキテクチャの中で検討するものはなかった．本研究では，ACT-R を用いて内発的動機づけの認知モデルの構築を目指す．モデルは環境中のパターンの発見を知的好奇心の源泉とみなす．それによって，モデルは複数の異なる広さの環境を学習することができた．

**キーワード：認知モデリング，ACT-R，内発的動機づけ**

## 1. はじめに

人間は，外部から与えられる報酬だけでなく，内部から発生する報酬を利用し，目的を達成するために様々な環境を学習することができる．このような人工エージェントの意欲的な学習を実現するために，自己効力感や好奇心などの報酬によって駆動される内発的動機づけの概念が，複数の研究者によって議論されてきた [15, 10].

しかし，強化学習の枠組みに基づいたこれらの研究は，内発的動機づけと他のプリミティブな認知機能との関連性を説明していない．これに対し，最近の認知モデルの研究では，様々なタスクで共通に利用される認知のプリミティブな機能を統合した認知アーキテクチャへの注目が高まってきている．異なるタスク間でプリミティブなプロセスを共有することで，人間の認知の全体的な構造を引き出すことができる．

このような研究に従って，本研究では，内発的動機づけの背後に存在する認知メカニズムの一つを，認知アーキテクチャに基づいて表現することを試みる．本研究では，複数存在する認知アーキテクチャのうち，ACT-R (Adaptive Control of Thought-Rational[1]) を用いる．このアーキテクチャは広く利用されており，先行研究も豊富である [7]．また，ACT-R は，脳の部位に対応し心理学実験で検証されているモジュールを複数有している．その中には，従来の自律エージェントで用いられている強化学習に近いモジュールも存在する．したがって，我々は数あるアーキテクチャの中から ACT-R を利用することが適していると考えた．本研究においては，ACT-R のプリミティブな認知機能を集積することで内発的動機づけをモデル化する．

このモデルを提案する前に，このテーマに関連する先行研究を示すことで目的を明確にする．その上で，人間の知的好奇心との対応を想定し，特に環境と内部知識とのパターンマッチに着目した内発的動機づけのメカニズムを提案する．提案したメカニズムは，特定のタスクのシミュレーションを行うために実装されている．最後に，現状をまとめ，今後の方向性を示す．

## 2. 関連研究

本節では，環境学習に関する研究の方向性として，強化学習に基づく研究と ACT-R に基づく研究を紹介する．

### 2.1 強化学習における内発的動機づけ

今日，内発的動機づけに基づく人工エージェントの研究が存在する [15, 10]．これらの研究では，内発的動機の一つである好奇心をモデル化し，エージェントに広く環境を探索させる方法を検討している．このような研究の多くは，強化学習などの統計的学習の枠組みを用いている．

通常，強化学習におけるエージェントは，外部環境から受け取った情報に基づき行動を決定する．行動の結果によって環境は報酬を発生させ，エージェントは時間経過の中で，それを最大化するように努める．この従来的な強化学習の枠組みに対し，Sutton はエージェントと環境間の境界と，身体と環境間の物理的

な境界は同様ではないとのべる [16]．この主張を援用し，Singh は，従来の強化学習に内発的動機を取り入れた Intrinsically motivated reinforcement learning (IMRL) [15] を提案した．外部環境から直接報酬を受ける従来の強化学習に対し，IMRL は，内部環境の状態によって報酬が変動する．これにより，予期せぬ反応に対する好奇心などをモデル化する．

近年，この内発的動機づけに基づく強化学習の研究は，深層強化学習の枠組みで発展している [2, 13]．特に，Burda は，エージェントの内側から得られる内因的報酬のみに基づいた環境学習を検討した [2]．Atari や Unity の迷路課題などのゲーム画面を入力 [11] とし，エージェントにとって未知の状況から内在的報酬を生成した．その結果，エージェントは幅広く環境を学習し，ゲームのスコアを向上させた．Burda は，エージェントのゲームスコアがあがった理由を，通常ゲーム環境はユーザーの好奇心を刺激するために設計されていることが多く，環境の中で新しい情報を見つけるとゲームの点数が上がるように設計されているからだと主張している．

### 2.2 ACT-R による環境探索と感情のモデル

前節で述べた関連研究は，最適な探索を実現する学習アルゴリズムの提案を目的としていた．しかし，人間とモデルの対応関係については検討されていない．

前節の先行研究に対し，ACT-R を用いることで，人間とモデルの対応関係について検討することができる．ACT-R は脳の部位に対応するモジュールを保持する認知アーキテクチャである．例えば，宣言的モジュールは経験や知識を保持し，ゴールモジュールはタスクの状態を管理する．ACT-R のプロダクションルールは，このようなモジュールの状態に基づいて選択され，アクションとしてモジュールにコマンドを送る（条件を満たす知識を探す，タスクの現在の状態を更新するなど）．これらのルールには，モジュールの状態と柔軟な対応（パターンマッチ）を実現する変数が含まれている．

環境学習について，Fu は，上下左右の方向に関する知識を実装し，繰り返し迷路課題を解くために ACT-R を用いた [19]．Fu のモデルは，Q-Learning[20] と同様の ACT-R の「ユーティリティ」モジュールを用いる．そのモデルでは，現在の目標達成につながる行動をしたときには正の報酬を，目標達成につながらない行動をしたときには負の報酬を受け取るようにした．このような報酬を得る試行を蓄積することで，モデルは最適な行動をとるように学習した．

他の研究では，迷路課題を，強化学習だけでなく，ACT-R で実装された宣言的知識の学習（事例ベース学習）を用いて行った．Reitter は，モデルの宣言的知識に位置情報を入れ，トポロジカルマップを構成し，それを利用したバックトラッキングを用いて経路探索を行うモデルを構築した [14]．

ACT-R による環境探索に内発的動機づけを組み入れる直接的な研究は存在しないものの，内発的動機づけと密接に関係する感情的要素のモデルは多く提案されている．Dancy は ACT-R による認知プロセスと生理的なメカニズムとの結合によって感情の生起を説明した [3]．また，Vugt は，抑うつを感情を伴う記憶の割合によって説明する認知モデルを構築した [18]．さらに，Juvina は，そのような感情的記憶と報酬関数の関係を説明することを目指した [5]．

これらの研究はいずれも，ACT-R の新しいモジュールや機能を開発し，感情プロセスにアプローチしている．これに対し，本研究では，ACT-R のプリミティブな機能を用いて，内発的動機づけをモデル化することを目的としている．特に本研究では，ACT-R の学習過程から自然に生まれた報酬変動のメカニズムを提案する．

## 3. 内発的動機づけのメカニズムの提案

本節では，我々が提案する内発的動機づけのメカニズムを提案する．このメカニズムは，知的好奇心とパターンマッチを対応させるという考えに基づいている．この考えを説明した後，ACT-R のプリミティブな機能を集積させた内発的動機づけの一般的なフレームワークを示す．

### 3.1 知的好奇心と ACT-R モデルの対応

本研究では，Burda に従い，内発的動機づけの要因として好奇心に注目する．先行研究に示されるように，好奇心は環境探索を促進し，ゲームにおけるパフォーマンスを向上させる．実際，ゲームデザイナーである Koster は，著書の “Theory of Fun for Game Design [6]” において，優れたゲームがユーザの好奇心を刺激するものであることを述べている．特に，Koster は，人間が感じる楽しさは，環境や状況において，新しいパターンを発見することによって引き起こされると述べる．例えば，数あるパターンから最適解が発見されたゲームは，「飽き」が生じてしまう．

ここで，人間によるパターンの発見に対応する概念として，パターンマッチという仕組みに注目する．パターンマッチは汎用的な仕組みである．例えとして，テキスト検索で用いられる正規表現で表現されたパターンマッチがあげられる．ACT-R においてパターンマッチは，プロダクションルールとモジュールの状態のマッチングに利用される．図 1 は，ACT-R のパターンマッチを示している．この例では，ACT-R プロダクションルールの THEN 句に含まれる変数「Var1」と「Var2」を，宣言的知識の定数である 1 や 2 などの数値に束縛させている．

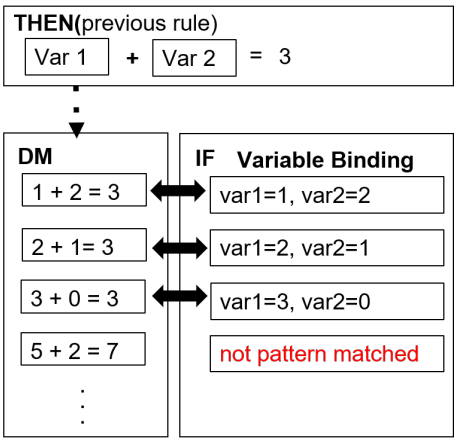


図 1　ACT-R のパターンマッチの簡単な例．この例では，前のルールの THEN 句で宣言知識のメモリを問い合わせ，次のルールの IF 句で変数が束縛される流れを示している．

このようにパターンマッチは，ルールに埋め込まれた変数のパターンに応じてデータや環境中の構造を発見する．この仕組みは人間の認知の柔軟性を説明するとされ，類推に代表される関係的推論，ゴールを状況に応じて変更するメタ認知など，人間に固有とされる認知機能のベースとなるともされる[1]．このことから，我々は，認知モデルのパターンマッチが，人間による知的探究（＝パターンの発見），すなわち知的好奇心と対応するのではないかと考えた．

## 3.2　知的好奇心の衰退

一般的な認知プロセスにおける知的好奇心の役割を説明するためには，そのような動機づけがタスク実行の過程でどのように減衰していくのかを考える必要がある．我々は，このような減衰のプロセスが飽きと対応すると仮定する．以下に，これを具体的に検討するために，ACT-R の学習機能である「ユーティリティ」と「プロダクションコンパイル」モジュールについて説明する．

### 3.2.1　ユーティリティ

ACT-R には，宣言的知識（チャンク）と手続き的知識（プロダクションルール）の 2 種類の知識がある．それぞれには，知識を獲得するための学習メカニズムと，知識の利用を変更するための学習メカニズムがある．本研究では，タスクを続けるか（楽しい状態），タスクをやめるか（飽きた状態）を決めるための手続き的知識の競合に着目する．ACT-R は，競合解決（特定の状況で発火可能なルールを選択すること）を制御し，得られた報酬によって効用値を更新するユーティリティモジュールを有している [19]．このモージュールから得られる報酬の量は，プロダクションコンパイルの働きによって変動する．

### 3.2.2　プロダクションコンパイル

「プロダクションコンパイル」とは，2 つの定義されているプロダクションルールを 1 つのプロダクションルールに統合する機能である [17]．ある課題に対しての一連のルールを反復発火することで，ルールの統合が起き，課題達成までに発火するルールの数が減る．通常，統合の対象となるルールは，条件節に変数を含むものとなる．つまり，ACT-R において，プロダクションコンパイルによって統合されるルールは，IF 節とモジュールの状態（たとえば宣言的モジュールに含まれる知識）の間でのパターンマッチを伴うものである．つまり，統合前のルールに含まれていた変数は，個別の宣言的知識の値によって置換される．そのため，定型的で自動的な課題遂行のルールが生成され，モデルの振る舞いは，人間の慣れと同様の振る舞いになる．

図 2 は後述のシミュレーションで使用する迷路課題の環境において，現在位置からゴール位置までの経路探索に関する ACT-R モデルのトレースを示す．縦軸に時間を示し，各列はモジュールのイベントを示している．左側のトレースは，宣言的知識を用いて経路探索を行うモデルの初期状態の処理を表している．右側のトレースはプロダクションコンパイルによる学習後，宣言的知識を取得せずに経路探索を行うモデルの処理を表している．

[1] Anderson は「ダイナミック・パターンマッチ」と呼ばれる ACT-R の機能の導入時にこの主張をした．

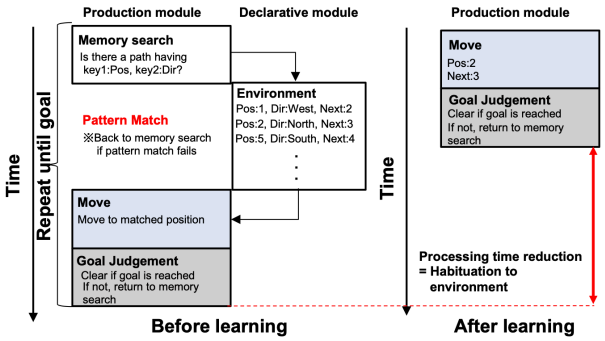


図 2　プロダクションコンパイルモジュールによる学習前と後の例

上記の仕組みを迷路環境の探索に当てはめれば，モデルはタスク初期において，環境地図の記憶検索を度々行う．タスクが進行するに従って，それらの記憶検索が必要なくなる．その結果，モデルによるタスクの遂行が効率化されるだけでなく，パターンマッチの頻度が減少する．つまり，環境におけるパターンが発見尽くされ，飽きをモデル化できると考える．

### 3.3　タスク継続のメカニズム

前節で提示した ACT-R のプリミティブな機能を用いて，我々はタスクを継続するか停止するかを判断する知的好奇心のメカニズムを提案する．図 3 は，我々が提案する一般的な環境におけるタスク継続のメカニズムである．各ラウンドの開始時に，モデルはタスクを継続するか停止するかを判断する（2 つのルールの競合解決）．タスクの継続を判断した後，モデルは様々なルール（マップの探索など）を発動させてラウンドを進める．ラウンドを終了する条件に遭遇すると，新たなラウンドを開始し，再びタスクを継続するか停止するかを判断する．

上記のプロセスにおいて，タスク続行ルールの効用値の初期値は，タスク終了ルールの効用値の初期値より高く設定されると考える．タスク開始時に，人間はある程度はタスク継続の意思があると考えるためである．この初期状態からの「飽き」のプロセスは，各ラウンドの終了を認識するルールの発火を負の報酬のトリガーとすることでモデル化できる．ラウンド終了時に負の報酬が与えられることで，そのラウンドにおいて競合解決の結果として発火したルール（すなわちタスク続行ルール）の効用値が低下し，タスク終了ルールが選択される確率が増加する．

「飽き」を抑止し，タスクを継続させる条件を検討するためには，タスクにおける「楽しさ」のモデルが

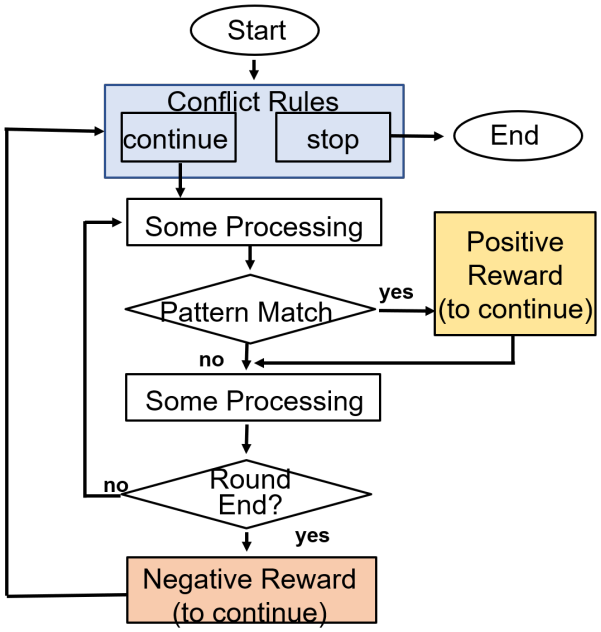


図 3　タスク継続モデルのフローチャート．パターンマッチが発生すると正の報酬が得られる．

必要である．タスク中に「楽しい」と感じるプロセスが生起した時に正の報酬がトリガリングされれば，タスク続行ルールの効用値は高い値を保ち，継続的なタスク継続が可能となる．中の宣言的知識の検索成功（地図を思い出せた，他者の振る舞いの予測できたなど）に付随して発火するルールと定義する．宣言的知識の検索はルールの IF 節（現在の状況）と宣言的知識内の記憶とのパターンマッチであり，これに成功することは Koster による楽しさの定義と整合する．ただし，このルールは繰り返し実行されると慣れ，つまりルールの統合が起こる．統合が起こるとルーティン化してしまい報酬を得られなくなり，タスク続行ルールの効用値が減少してタスク終了ルールが発火するようになる．つまりこのモデルにおけるタスク続行の要因は，パターンマッチの対象となる宣言的知識を見つけ続けられることとなる．

## 4.　実装

本研究の目的は，ACT-R のプリミティブな機能を用いて内発的動機をモデル化することである．この目的を達するためには，前節で示したモデルを具体的な課題の上に実装し，その振る舞いを観察する必要がある．本研究では，従来の内発的動機づけに関わる研究でも用いられてきた迷路探索を課題とする．

### 4.1 迷路課題のモデル

本研究の迷路探索のための ACT-R モデルの実装では，Reitter のモデル [14] をタスク継続のメカニズムを含むように拡張した（図 3）.

#### 4.1.1 迷路環境

迷路探索に対する ACT-R モデルの実装は，Reitter に従い，確率的な DFS (depth-first search) によってヒューリスティックに環境探索を行う．モデルは，各曲角の繋がりをノードとするトポロジカルマップを宣言的モジュールに保持する．トポロジカルマップはノードとノードを結合するチャンクの集合として表現され，モデルはそれを検索することで環境探索を行う．迷路環境は，縦横 5×5, $7 \times 7$, $9 \times 9$ の大きさのマップを各々 10 マップずつ用意した．これらマップの初期位置は迷路の最も左上の曲角であり，ゴール位置はスタート位置から最も遠い曲角である．つまり，スタート位置から経由するホップ数が最も多い曲角がゴール位置として選択される．なお，複数のゴール位置の候補があった場合は，候補の中から最も早くホップ数を計算したものが決定される．

#### 4.1.2 環境探索

それぞれの環境において，モデルは現在自分が位置する曲角をゴールバッファに格納する．モデルの課題は，スタートからゴールにゴールバッファ内の現在位置を遷移する事である．現在位置の遷移は，宣言的モジュールに格納されたチャンクを検索することで行われる．現在位置と結合するチャンクが呼び出され，そのチャンクと結びつく位置が新たにゴールバッファに格納される．これをゴールに達するまで繰り返す．

モデルは，ゴールをするたびに，現在地からゴールまでに想起されたチャンクに対して，それを正解とするラベルを付与し，宣言的記憶に格納する．次のラウンドからは，モデルはこれらのラベル付けされたチャンクを用いて，事例ベース学習理論の手法に従い，タスクを効率的に実行する [8].

モデルがラベル付けされたチャンクを取得できなかった場合，モデルは現在位置からゴール位置までのパスを，確率的な深さ優先探索 (DFS) というヒューリスティックな探索を用いて経路探索する．DFS で用いられるバックトラッキングを実現するために，ACT-R のイマジナルモジュールを用いてスタック構造を実装した．図 4 は，このモジュールで生成されたチャンクを用いたスタック構造を示している．スタック内のプッシュ機能は，ARG1 スロットに過去のチャンク名を格納するチャンクを生成することで実現している．また，スタック内のポップ機能は，ARG1 スロットの値を過去のスロットの値に戻すことで実現している．これらの生成されたチャンクは宣言的知識に格納され，後から取得してポップ機能を実現することができる．これらの処理はすべて，LISP など他のプログラミング言語で書かれた外部関数を定義することなく，ACT-R の生成ルールのみで実装した．

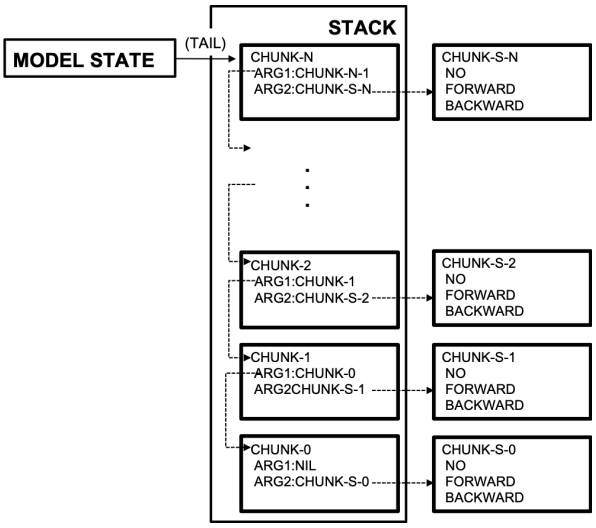


図 4　ACT-R のチャンクで構築したスタック構造．スタックの実装のために ACT-R のイマジナルモジュールを使用している．

#### 4.1.3 タスクにおいての内発的動機づけ

この単純な迷路課題の中で，前節で示した楽しさや飽きのプロセスがどのように生じるのかを検討した．このモデルにおいて，パターンマッチに付随する「楽しさ」は，現在の状況から宣言的知識に記憶しているパスを思い出すことと定める（パスの発見時に発火するルールを正の報酬のトリガーとする）．また，パス探索の終了を課題の継続のモチベーションの減少と定義する．つまり，ゴール探索の成功，失敗によらず，ラウンドの終了時に発火するルールを負の報酬のトリガーとする．課題を継続すると継続ルールの効用値は，負の報酬によって減少していく．そして，継続ルールの効用値が，終了ルールの効用値を下回った時に，終了ルールが発火し，課題が終了することになる．

## 4.2 シミュレーション

### 4.2.1 設定

実装されたモデルによる内発的動機づけの挙動を確認するシミュレーションを実施した．本シミュレーションの環境は，我々がこれまでに行ったシミュレーションの環境を拡張したものである [12]．環境は，広さを変数に持ちマップの広さを変化させることができる．加えて，各々の広さの中で，モデルは複数の異なるマップにおいて同様の課題を行うことができる．

この環境において，内発的動機づけの効果をはかる指標として，モデルがマップの曲角を訪れる回数をもとにしたエントロピーを用いる．エントロピーは不確実性をはかる指標である．本シミュレーションにおいて，モデルが幅広く環境を探索，つまりモデルが多くの曲角に訪れていればエントロピーが増加する．したがって，エントロピーは，モデルが幅広く環境を探索しているかどうかの指標になるのではないかと考えた．これらの環境において，継続ルールの効用値の初期値を 10，終了ルールの効用値の初期値を 5 とした[2]．また，ラウンドの終了ルールには負の報酬のトリガー $(r = 0)$ を，パターンマッチを含むルールには正の報酬のトリガーを割り当てた．加えてパスの発見によってトリガーされる報酬値を 1 から 20 まで変化させ，各報酬値に対して 1000 回，課題の継続シミュレーションを行った．なお，各ラウンドの制限時間をそれぞれ 100 秒と設定した．制限時間に達するとモデルはゴールの達成によらず次のラウンドに移行した．

### 4.2.2 結果

図 5 は，縦横 5×5, $7 \times 7$, $9 \times 9$ の環境において，パスの発見に伴う報酬値を変化させた際のシミュレーション結果を示す．それぞれのグラフは，ラウンド継続数，コンパイルモジュールによって生成されるルールの数，エージェントが環境の曲角を訪問した回数のエントロピー，ゴール達成率の変化をグラフ化したものである．報酬値が大きいほどラウンド継続数，ルール数，エントロピーが大きくなっていく．これに伴い，ゴール達成率が大きくなっていっている．また，マップの広さを変更した際も同様の傾向が見られる．これらの結果からエントロピーの増加によって，モデルは幅広く環境を探索する事を確認できた．

以上のことより，我々の提案した内発的動機づけのメカニズムがよく機能したことが示される．加えて，同様の広さのマップでも，特にゴール達成率に差異が生じている事が確認できる．マップの広さに着目すると，マップが広ければ広いほど全体的にゴール達成率が低下する．また，マップが広いほどルール数は全体的に増加するが，その増加数は微々たるものである．モデルの学習は飽和状態になっていると考える．人間は，複雑過ぎる環境において，学習における内発的動機が低下する [9] つまり，モデルは，マップの広さと 100 秒という制限時間において，学習の限界を向かえていたと考える．

## 5. 既存手法との対比

我々が提案した ACT-R のモデルの振る舞いを明確化するために，同様の環境で動作する強化学習モデルを実装し，シミュレーションを行った．実装したモデルは，IMRL を援用した Q-Learning を行う．方策には $\varepsilon$-グリーディ法[3]を用いた．モデルが 1 ラウンドに行動できる限界を 100 ステップとした．この行動限界までに，モデルは，スタートからゴールまでたどり着けば，課題クリアとなる．式 1 は，本モデルの Q 値の更新式である．$r_e$ は外部の環境からの報酬を $r_i$ は内部の報酬を表す．モデルは外因的報酬，つまり $r_e$ を，道を思い出せなかった場合は負の報酬 (-1)，移動できた場合は 0 として，ゴールのできた場合は正の報酬 (10) を受け取る．

$$Q(s,a) \leftarrow Q(s,a) + \alpha[r_i' + r_e' + \gamma max Q(\acute{s},\acute{a}) - Q(s,a)] \quad (1)$$

式 2 の $p$ は Q 値に対しての遷移確率を表す．確率 $p$ に対して，余事象の確率のエントロピーを $r_i$ としている．$r_i$ は 1 ステップ毎のモデルの内発的動機である．また，$\tau$ はシミュレーション時に用いる報酬値のための係数である．この係数を 0.35 から 0.73 まで 20 段階変化させ各報酬値に対して 10000 回シミュレーションを行った．1 ラウンド中の内因的報酬の合計値が，5 以下であればラウンドの終了とした．

$$r_i = -\tau(1-p)\log(1-p) \quad (2)$$

図 6 は，$\tau$ を変化させた際に，ラウンド継続数，エージェントが環境の曲角を訪問した回数のエントロピー，ゴール達成率がどのように変化するのかをグラフ化したものである．ACT-R モデルの結果に対し，これらの結果は，同じ広さのマップ間によって差異が目立ち，安定してないように思われる．また，内因的報酬に関わる $\tau$ の値が大きいほど，モデルは課題を継続する

[2]探索方向（上下左右）を確率的に決める各ルールの効用値の初期値をそれぞれの 10 とした．記憶の想起のノイズである ans (activation noise level) = 0.1，継続ルールと終了ルールの効用値の比較のノイズである egs (expected gain s) = 1 と設定した．

[3]$epsilon = 0.2$, $gamma = 0.9$, $alpha = 0.2$

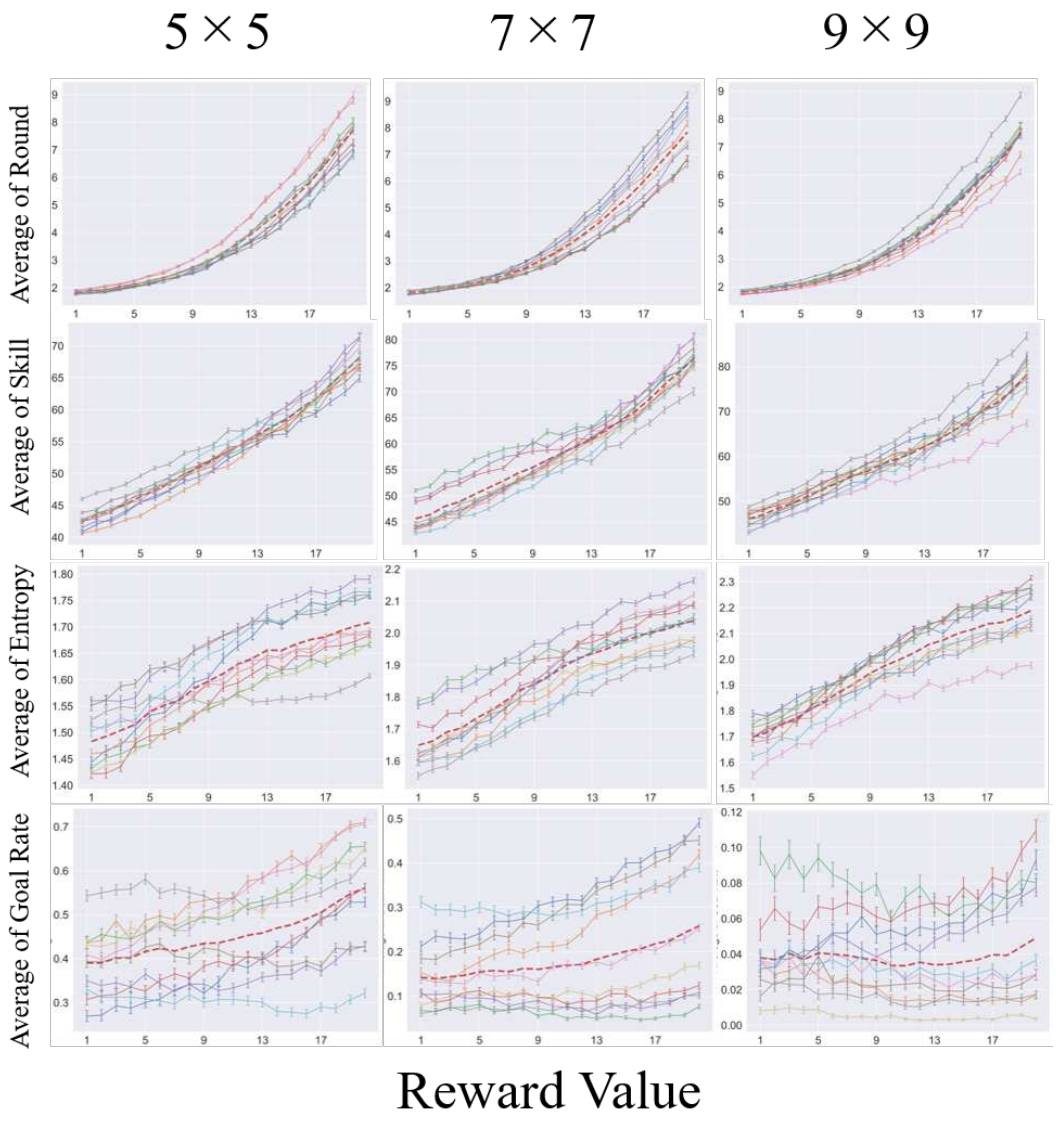


図 5　5 × 5, 7 × 7, 9 × 9 マップの ACT-R モデルの内部報酬値におけるラウンド継続数，ルール生成数，エントロピー，ゴール達成率の推移．エラーバーは標準誤差を表す．赤のダッシュ線はグラフの平均である．

が，ゴール達成率は下がっている．加えて，エントロピーが増加する環境や減少するマップが存在する．

このモデルの詳細な行動を確認するため，それぞれの $\tau$ の値による 9 × 9 の大きさの時のエントロピーの推移を詳細に確認した．図 7 は，図 6 のうち，エントロピーが減少しているマップと増大しているマップを，それぞれ 1 つずつ抜き出したものである．図より $\tau$ の値の変化による減少，増大の傾向を確認できる．また，この時のマップの曲角の訪れた頻度をヒートマップとして図 8 に示す．上段がエントロピーが減少している時の環境，下段がエントロピーが増大している時の環境である．黄色と緑のひし形は，それぞれの環境においてのスタートとゴールを示す．各々の曲角の色の濃さは，モデルが訪れた回数に比例して濃くなる．また，各曲角に出力されている数字は，各曲角の訪れた回数を環境全体の曲角の訪れた回数で正規化したものである．どちらも $\tau$ が大きくなるにつれて，スタート付近の曲角の色の濃さが薄くなり，探索しているように思える．そのため，先行研究にもあるように内発的動機づけのパラメータの増大によって環境を探索する幅が広がっていると思われる．しかし，このモデルの動作はゴール達成率に寄与していない．おそらく，モデルは，スタート付近の曲角を中心とした移動を繰り返していると思われる．そのため，左のグラフは，スタート付近の曲角が行き止まりになっているため，行き止まりからスタートに戻るという動作を繰り返す回数が

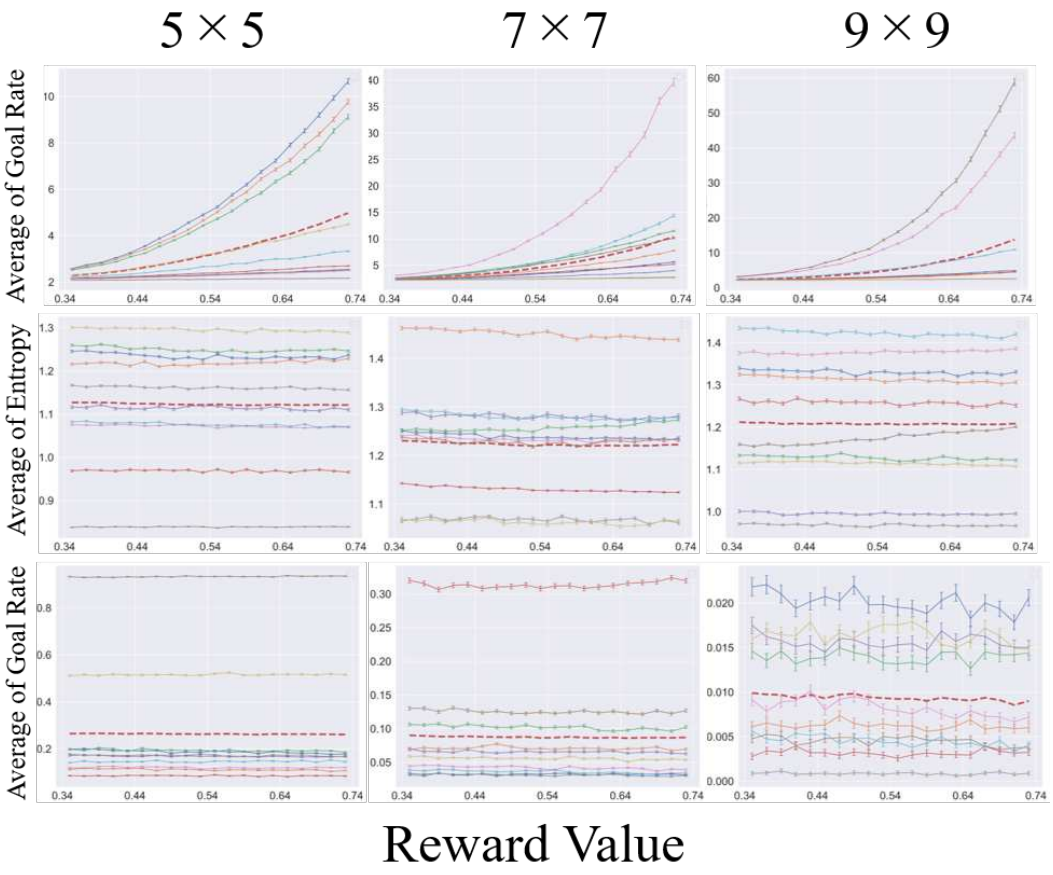


図 6　5 × 5, 7 × 7, 9 × 9 マップの強化学習モデルの内部報酬値におけるラウンド継続数，エントロピー，ゴール達成率の推移．エラーバーは標準誤差を表す．赤のダッシュ線はグラフの平均である．

増えることでエントロピーが減少しているのではないかと考える．

これらの結果から，エントロピーの増減がゴール達成率の増減に関連していないことが示される．一般的に，課題中に特別なイベントがない環境では（正の報酬はゴール時のみ），強化学習はうまく動作しない．そのため，モデルはこれらの環境を学習できていないと考えられる．このモデルの行動は，式の $r_i$ と $r_e$ のバランス調整や，Burda が指摘しているように，好奇心を刺激するために迷路の環境を注意深く設計することで変えられると思われる [2]．また，通常の強化学習に内発的動機づけを付与するだけでなく，特に IMRL で用いられていたオプションによるスキルの階層型の学習が必要であると思われる [15]．これは，ACT-R のコンパイルモジュールと同様，エージェントの環境においての慣れを表現し，エージェントが環境の探索済み箇所よりも未探索の箇所を探索する機会が増えることで，幅広く環境を探索するように振る舞うのではないかと考える．しかし，我々の ACT-R モデルは，このような慎重なパラメータ調整や環境設計，階層型学習の仕組みを実装しなくても，汎用的なモジュールを用いれば環境学習を行うことが可能である．このように，内発的動機づけを表現する上で，本モデルの優位性を主張することができた．

## 6.　まとめ

本研究の目的は，ACT-R と強化学習モデルの対応を検討するものであった．まず，ACT-R から提供されるプリミティブな認知プロセスの集積によって，認知モデルに内発的動機づけのメカニズムを実装した．この実装を行うために，Koster の理論を基に，パターンマッチのメカニズムが知的好奇心の源泉である「楽しさ」を表していると仮定した．その結果，パターンマッチが成功することで，課題継続に対する高い内発的動機づけが維持された．一方，コンパイル機能を用いてパターンマッチをスキップすることで，モデルは課題に対する「飽き」を生じさせ，課題を終了すると考えた．

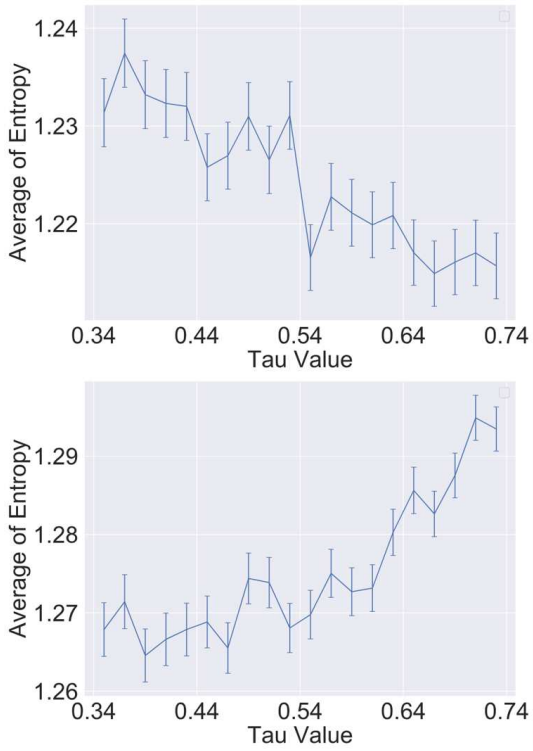


図 7　エントロピーが減少しているマップと増大しているマップの $\tau$ による推移.

この内発的動機づけのメカニズムの振る舞いを確認するため ACT-R のモデル実装し，シミュレーションを実施した．モデルは報酬値が大きいほどラウンド継続数，ルール数，エントロピーが大きくなった．これに伴い，ゴール達成率が上昇した．また，マップの広さを変更した際も同様の傾向が見られた．エントロピーの増加によって，モデルは幅広く環境を探索する事を確認できた．また，マップの広さと制限時間によって，モデルが学習の限界を向かえる振る舞いを確認できた．

この ACT-R のモデルに対し，IMRL を援用し，内発的動機づけを付与した強化学習モデルを実装し同マップにおいてシミュレーションを実施した．その結果は，同じ広さのマップ間によって差異が目立ち安定していなかった．また，内因的報酬に関わる $\tau$ の値が大きいほど，モデルは課題を継続するが，ゴール達成率は下がった．加えて，エントロピーが増加する環境や減少するマップが存在した．その後，エントロピーが減少したマップと増加したマップを取り上げ，モデルの詳細な振る舞いを観察した．エントロピーの増減に関わらずモデルのゴール達成率は減少していたためエントロピーの増減がゴール達成率の増減に関連していないことが示された．

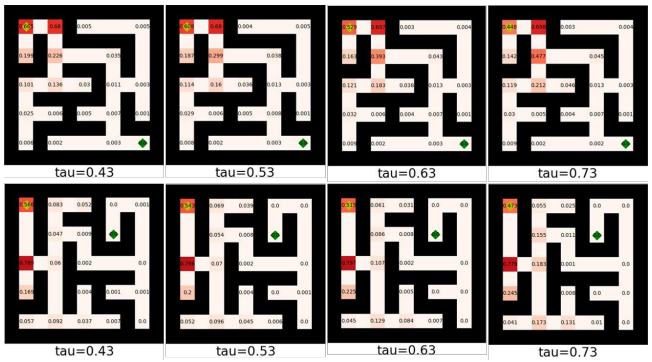


図 8　図 7 に対応するエントロピーが増大しているマップと，減少しているマップのヒートマップ.

これらのシミュレーション結果から，本モデルは新しい環境を学習する上で優位性を持っていると考えられる．このモデルでは，ユーティリティモジュールやコンパイルモジュールだけでなく，ゴールまでの正しい経路を記憶する事例ベースの学習 [8] を用いている．このようないくつかの学習アルゴリズムの組み合わせは，現在の迷路タスクにおける内在的報酬と外在的報酬のバランスをとるのに役立つかもしれない．

また，我々のモデルの内在的報酬の表現は，他の先行研究と比較しても新規性がある．我々のモデルでは，内発的報酬と外発的報酬を式の中で明示的に分けてはいないが，内発的動機づけの効果は既存の ACT-R のメカニズムの中で自然に表現されている．このアプローチは，人間の認知理論に基づいていること，既存の学習研究との関連性があること，理論上の不要な要素を省けることなどの利点があると考えている．

今後の研究では，内発的動機づけのモデルを人間のデータと比較する必要がある．人間の認知のモデルとして，従来の強化学習で提示された行動が誤りではなかったかもしれない．探索作業中，人はしばしば目標を忘れてパフォーマンスを低下させる．このような非合理的な行動は,「計算精神医学」のテーマにも関連しているかもしれない [4].

今後の課題として，動機づけの最適水準 [21] に達するまでのモデル化が必要である．本モデルは静的に継続ルールの効用値を決めている．つまり，最適水準に達するまでの過程はモデル化されていない．したがって，モデルが最適水準に達するまでの課題の検討が必要になってくるだろう．

## 文献

[1] J. R. Anderson. *How Can the Human Mind Occur in the Physical Universe.* Oxford Press, 2007.

[2] Yuri Burda, Harri Edwards, Deepak Pathak, Amos Storkey, Trevor Darrell, and Alexei A Efros. Large-scale study of curiosity-driven learning. *arXiv preprint arXiv:1808.04355*, 2018.

[3] C. L. Dancy, F. E. Ritter, K. A. Berry, and L. C. Klein. Using a cognitive architecture with a physiological substrate to represent effects of a psychological stressor on cognition. *Computational and Mathematical Organization Theory*, Vol. 21, No. 1, pp. 90–114, 2015.

[4] Quentin JM Huys, Tiago V Maia, and Michael J Frank. Computational psychiatry as a bridge from neuroscience to clinical applications. *Nature neuroscience*, Vol. 19, No. 3, p. 404, 2016.

[5] Ion Juvina, Othalia Larue, and Alexander Hough. Modeling valuation and core affect in a cognitive architecture: The impact of valence and arousal on memory and decision-making. *Cognitive Systems Research*, Vol. 48, pp. 4 – 24, 2018. Cognitive Architectures for Artificial Minds.

[6] R. Koster. *Theory of Fun for Game Design*. ParaglyphPr, 12 2004.

[7] Iuliia Kotseruba and John K. Tsotsos. 40 years of cognitive architectures: core cognitive abilities and practical applications. *Artificial Intelligence Review*, Jul 2018.

[8] Christian Lebiere, Cleotilde Gonzalez, and Michael Martin. Instance-based decision making model of repeated binary choice. 2007.

[9] Thomas W Malone. Toward a theory of intrinsically motivating instruction. *Cognitive science*, Vol. 5, No. 4, pp. 333–369, 1981.

[10] A. Manoury, M. N. Sao, and B. Cédric. Hierarchical affordance discovery using intrinsic motivation. In Proceedings of the 7th International Conference on Human-Agent Interaction (HAI ' 19), pp. 186–193.

[11] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A Rusu, Joel Veness, Marc G Bellemare, Alex Graves, Martin Riedmiller, Andreas K Fidjeland, Georg Ostrovski, et al. Human-level control through deep reinforcement learning. *Nature*, Vol. 518, No. 7540, pp. 529–533, 2015.

[12] K. Nagashima, J. Morita, and Y. Takeuchi. Modeling intrinsic motivation in act-r: Focusing on the relation between pattern matching and intellectual curiosity. In *ICCM2020: 18TH INTERNATIONAL CONFERENCE ON COGNITIVE MODELING*.

[13] Deepak Pathak, Pulkit Agrawal, Alexei A Efros, and Trevor Darrell. Curiosity-driven exploration by self-supervised prediction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pp. 16–17, 2017.

[14] D. Reitter and C. Lebiere. A cognitive model of spatial path-planning. *Computational and Mathematical Organization Theory*, Vol. 16, No. 3, pp. 220–245, 2010.

[15] S. Singh, A. G. Barto, and N. Chentanez. Intrinsically motivated reinforcement learning. In L. K. Saul, Y. Weiss, and L. Bottou, editors, *Advances in Neural Information Processing Systems 17*, pp. 1281–1288. MIT Press, 2005.

[16] R. S. Sutton and A. G. Barto. *Reinforcement Learning: An Introduction*. MIT Press, 2 1998.

[17] Niels A Taatgen and Frank J Lee. Production compilation: A simple mechanism to model complex skill acquisition. *Human Factors*, Vol. 45, No. 1, pp. 61–76, 2003.

[18] Mj .K van Vugt and M van der Velde. How does rumination impact cognition? a first mechanistic model. *Topics in Cognitive Science*, Vol. 10, No. 1, pp. 175–191, 2018.

[19] F. Wai-Tat and J. R. Anderson. From recurrent choice to skill learning: A reinforcement-learning model. *Journal of experimental psychology. General*, Vol. 135, pp. 184–206, 6 2006.

[20] C. J. C. H. Watkins. *Learning from delayed rewards*. King's College, Cambridge, 1989.

[21] Robert M Yerkes and John D Dodson. The relation of strength of stimulus to rapidity of habit-formation. *Journal of comparative neurology and psychology*, Vol. 18, No. 5, pp. 459–482, 1908.