3　内発的動機づけの ACT-R モデル

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 [Koster 04]” 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．ACT-R において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．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，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 [Anderson 07]．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）の楽しみと対応するのではないかと考えた．この考えから，本研究では ACT-R のもつパターンマッチ，およびそれと関連する汎用的な認知機能を集積することで「楽しさ」をモデル化する．

3.2　ACT-R の学習理論

ACT-R は複数の汎用的なモジュールを持っている．パターンマッチとして表現される「楽しさ」に対して，「飽き」を表現するために「プロダクションコンパイルモジュール」と「ユーティリティモジュール」に着目する．以下，これらの機能の概略を説明する．

3.2.1　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 [Anderson 86]．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．統合の対象となるルールは，IF 節に変数が含まれ，モジュールの状態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手続きがシミュレーションされ，人間的な慣れと同様の振る舞いをする [Anderson 87]．

3.2.2　ユーティリティ

ACT-R における効用（utility）とは，複数のルールの競合解消に利用されるパラメータであり，個別のルールに付与される．ACT-R の「ユーティリティモジュール」は，効用値の更新を制御するモジュールであり，効用関数の計算に報酬を用いる．図 1 は一般的なゲーム課題における環境探索の継続を示したものである．ゲーム内の各ラウンドの開始時において，ゲーム続行，ゲーム終了の判断（競合解消）を行う．ゲーム続行が選択された後，新たなラウンドがスタートする．

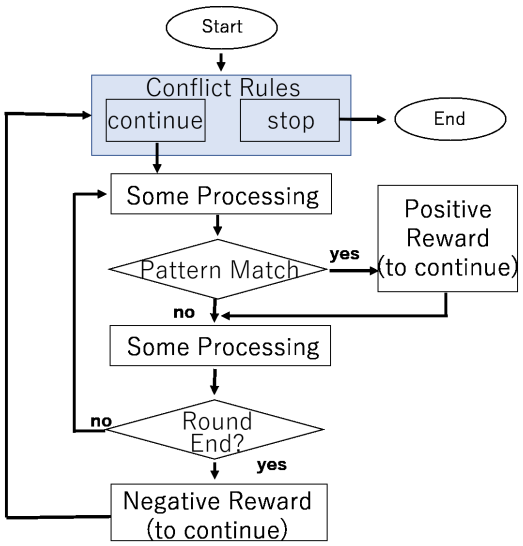

図 1: 課題継続モデルのフローチャート

上記のプロセスにおいて，ゲーム開始時に，人間はある程度はゲーム継続の意思があると考えるため，ゲーム続行ルールの効用値の初期値は，ゲーム終了ルールの効用値の初期値より高くされると考える．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ゲーム続行ルールに与えられることで，ゲーム終了ルールが選択される確率が増加する．

「飽き」を抑止し，ゲームを継続させる条件を検討するためには，ゲームにおける「楽しさ」のモデルが必要である．ゲーム中に「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，ゲーム続行ルールの効用値は高い値を保ち，継続的な課題継続が可能になる．本研究では，正の報酬のトリガーとなるルールを，ゲーム中の宣言的知識の検索成功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ングであり，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．

4　実装

4.1　概要

本研究は，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ものである．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．

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．また，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．初期位置は 16 であり，それを 1 まで遷移させることが課題となる．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する．次回以降のゲームにおいて，正解ラベルの付与されたチャンクを検索することで効率的にゴールに到達することができる．しかし，ゴールに繋がるチャンクを検索できなかった場合，記憶中のトポロジカルマップを辿りつつ，環境探索を行う．

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．また，パス探索の終了を課題継続のモチベーションの減少と定義する．

4.2　シミュレーション

実装されたモデルによる内発的動機の挙動を確認するシミュレーションを実施した．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．また，探索方向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．その他パラメータは ans (activation noise level) = 0.1， egs (expected gain s) = 1 を設定した．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 秒から 300 秒に設定した．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．

図 2 は，パスの発見に伴う報酬値を変化させた際に，ゴール達成率，課題継続数，コンパイルモジュールによって生成されるルールの数がどのように変化するのかをグラフ化したものである．報酬値が大きいほど継続数が大きくなっていく．これに伴い，ゴール達成率および，獲得されるルールの数も多くなっていっている．このことから，モデルによる環境学習において，設定され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．

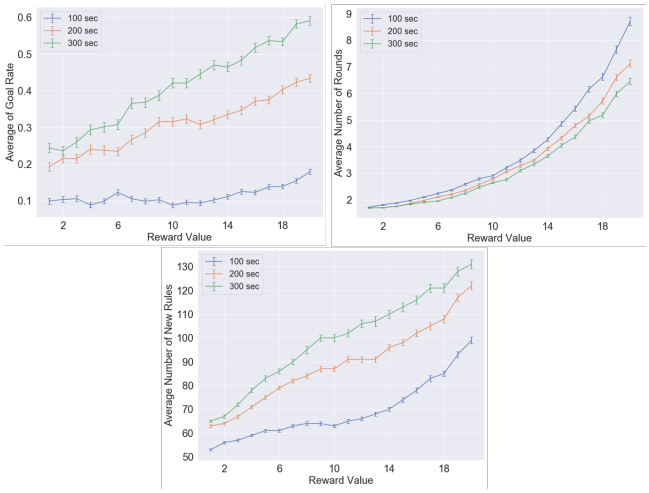

図 2: ACT-R モデルの内部報酬値におけるゴール達成率，ラウンド継続数，ルール生成数の推移 (SE)

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 ε -グリーディ法 ($\epsilon = 0.2$) を用い，ACT-R モデルと同様に各時点においてトポロジカルマップ中の曲角の一つを状態とし，西，北，東，南の 4 つの方向へ移動するモデルを実装した．他の固定値は，学習割引率 ($\gamma = 0.9$)，学習率 ($\alpha = 0.1$) とした．モデルが 1 ラウンドに行動できる限界を 100，125，150 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-1，移動できた場合 0，ゴール時できた場合 10 を受け取る．

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

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．我々は，1 ラウンド間のこの計算の合計が，ACT-R モデルの継続ルールの効用値と対応づけた．また，τ 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 1000 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を用いて，閾値 (5) 以下であればラウンドの終了とした．

$$r_i = -\tau(1-p)\log(1-p) \tag{2}$$

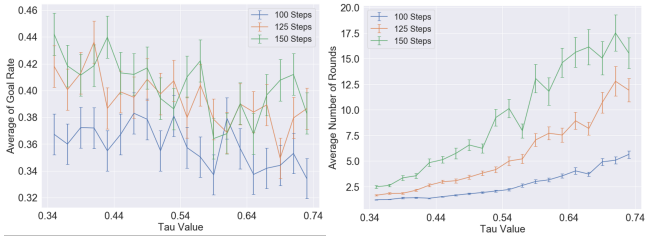

図 3: 強化学習モデルの内部報酬値におけるゴール達成率とラウンド継続数の推移 (SE)

図 3 は，τ を変化させた際に，ゴール達成率，課題継続数の変化をグラフ化したものである．この結果は，ACT-R モデルの結果に比べ分散が目立ち，安定してないように思われる．また，内因的報酬に関わる τ の値が大きいほど，モデルは課題を継続するが，ゴール達成率が下がっている．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，モデルはこの環境を学習できていないと考えられる．

また，ACT-R はモデルの内部時間もシミュレーションできる。そのため，環境における「時間」を表現する事は容易であった．しかし，強化学習において，環境における「時間」を適切に設定することは難しい．我々は ACT-R の課題の制限時間を，強化学習モデルのステップ数で表現しようと試みた．ACT-R の課題では制限時間を大きく設定すると，モデルの課題継続数は下がっていく．一方，強化学習モデルは，制限ステップ数を大きく設定すると，モデルの課題継続数は上がっていく．ACT-R のモデルの振る舞いは，1 ラウンドに対して探索できる時間が長いので，環境におけるパターン（「知的好奇心」）が，少ない課題継続の段階で取りつくしてしまい「飽き」につながったためだと考えられる．これに対して強化学習のモデルの振る舞いは対照的であった．この振る舞いは，モデルがゴールを達成するよりも，確率的に珍しい方向に向かうことを優先したためだと考えられる．同様の課題で内発的動機づけを表現できたが，モデルの好奇心を刺激する環境ではなかったと考える．

6 まとめ

本研究の目的は，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することであった．この目的の達成のために，Koster の理論を基に，人間のパターンの発見とモデルのパターンマッチが対応しているという仮定を置き，モデルが感じる「楽しさ」を，状況と宣言的記憶とのパターンマッチングの成功に対応付けた．課題環境に関わる先読みには，宣言的知識が利用されると仮定した．よって，これらのプロセスが成功することによって，モデルは課題継続の報酬をえる．そして，この利用がコンパイルによってスキップされることで，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．

提案されたアルゴリズムは，静的な状況における迷路探索の課題に適用された．それを検証するためモデルを作成し，シミュレーションを行いモデルの振る舞いを確認した．得られたシミュレーション結果から，パターンマッチを伴うルールに対する報酬の付与が，課題継続のルールの効用値を上昇させ，長期間にわたる課題遂行，およびそれに伴う課題パフォーマンスの向上を導いた．さらに，課題の継続にともない，モデルはコンパイルによって多数のルールを生成した．この結果から，本研究において提案するパターンマッチによる楽しさのモデル化が，本課題における環境学習に有効に働いたことが示された．

ACT-R のこのようなモデル化は，従来の自律学習における内発的動機をよりプリミティブなレベルで説明するものである．また，ACT-R による過去の研究においてもパターンマッチによる報酬の付与や，コンパイルによるタスクへの飽きを表現するものはなく，本研究において示したモデルは，オリジナルのものといえる．

今後の課題として，動機づけの最適水準 [Yerkes 08] 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．

加えて，複数の環境においてシミュレーションを行う必要がある．本研究で実装した強化学習モデルにおいて，内因性の報酬に伴ってゴール達成率が低下したことは，本研究において設定した環境が原因であった可能性がある．先行研究 [Burda 18, Pathak 17] が指摘した通り，内発的動機づけのモデルにおいて高いパフォーマンスを達成するためには，好奇心を刺激する環境デザイン (適切な報酬値やイベントの配置) が必要

である．この事から，複数の環境をエージェントに探索させることで，モデルの好奇心を刺激する，つまりモデルが楽しいと感じる環境を検討できるのではないかと考える．

参考文献

[Anderson 86] Anderson, J.: Knowledge compilation: The general learning mechanism, *Machine learning: An artificial intelligence approach*, Vol. 2, pp. 289–310 (1986)

[Anderson 87] Anderson, J. R.: Skill acquisition: Compilation of weak-method problem situations., *Psychological review*, Vol. 94, No. 2, p. 192 (1987)

[Anderson 07] Anderson, J. R.: *How Can the Human Mind Occur in the Physical Universe*, Oxford Press (2007)

[Burda 18] Burda, Y., Edwards, H., Pathak, D., Storkey, A., Darrell, T., and Efros, A. A.: Large-scale study of curiosity-driven learning, *arXiv preprint arXiv:1808.04355* (2018)

[Dancy 15] Dancy, C. L., Ritter, F. E., Berry, K. A., and Klein, L. C.: 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)

[Juvina 18] Juvina, I., Larue, O., and Hough, A.: 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

[Koster 04] Koster, R.: *Theory of Fun for Game Design*, ParaglyphPr (2004)

[Little 03] Little, D.: Learner autonomy and second/foreign language learning, *Guide to Good Practice* (2003)

[Manoury] Manoury, A., Sao, M. N., and Cédric, B.: Hierarchical Affordance Discovery using Intrinsic Motivation, In Proceedings of the 7th International Conference on Human-Agent Interaction (HAI ' 19), pp. 186–193

[Mnih 15] Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., Graves, A., Riedmiller, M., Fidjeland, A. K., Ostrovski, G., et al.: Human-level control through deep reinforcement learning, *Nature*, Vol. 518, No. 7540, pp. 529–533 (2015)

[Newell 73] Newell, A.: Production systems: Models of control structures, in *Visual information processing*, pp. 463–526, Elsevier (1973)

[Pathak 17] Pathak, D., Agrawal, P., Efros, A. A., and Darrell, T.: Curiosity-driven exploration by self-supervised prediction, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pp. 16–17 (2017)

[Reitter 10] Reitter, D. and Lebiere, C.: A cognitive model of spatial path-planning, *Computational and Mathematical Organization Theory*, Vol. 16, No. 3, pp. 220–245 (2010)

[Singh 05] Singh, S., Barto, A. G., and Chentanez, N.: Intrinsically Motivated Reinforcement Learning, in Saul, L. K., Weiss, Y., and Bottou, L. eds., *Advances in Neural Information Processing Systems 17*, pp. 1281–1288, MIT Press (2005)

[Sutton 98] Sutton, R. S. and Barto, A. G.: *Reinforcement Learning: An Introduction*, MIT Press (1998)

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

[Wai-Tat 06] Wai-Tat, F. and Anderson, J. R.: From recurrent choice to skill learning: A reinforcement-learning model, *Journal of experimental psychology. General*, Vol. 135, pp. 184–206 (2006)

[Watkins 89] Watkins, C. J. C. H.: *Learning from delayed rewards*, King's College, Cambridge (1989)

[Yerkes 08] Yerkes, R. M. and Dodson, J. D.: 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)