相反する適応的最適化計算による非合理的なスリル行動の認知モデル
Cognitive Modeling of Thrill-Seeking Irrational Behavior with Conflicting Adaptive Optimizations

高田 亮介[†]，坂本 孝丈[‡]，竹内 勇剛[‡]
Ryosuke Takata, Takafumi Sakamoto, Yugo Takeuchi
[†] 東京大学，[‡] 静岡大学
University of Tokyo, Shizuoka University
takata@sacral.c.u-tokyo.ac.jp

概要

スリルを楽しむといった非合理的な遊び行動は，生物が生きていることを自覚するための重要な要素であると考えられている．本研究では，進化と学習という生物のプリミティブな適応によってスリルを楽しむという非合理的な遊び行動を生成する認知モデルを提案する．このモデルでは，進化計算（GA）によって獲得した “危険を避ける” という生得的な状態価値をベースに，相反する報酬関数を用いて強化学習（Q 学習）を行うことで “危険を冒す” という経験的な状態価値を実現する．シミュレーション実験により，相反する報酬関数とスリルを求める度合いが，スリルを楽しむという非合理的な認知過程をモデル化するうえで有効であることが示唆された．

キーワード：スリル, 進化計算, 強化学習

1. はじめに

遊園地で急速に落下するジェットコースターに乗ったり，鬼ごっこ遊びであえて鬼に近づき挑発するといったように，人はしばしばあえて危険を冒すリスク行動によってスリルを楽しむことができる．このような危険に向かうリスク行動は，生物進化において重要な “危険を避けて生き残る” という側面において非合理的である．カイヨワは，スリルを楽しむといった理知的に矛盾する非合理的な行動が遊びの本質であり，それこそが生物にとって生きることを自覚するための根源的な性質であると述べた [1]．

これまで，スリルを楽しむという人の行動プロセスを構成論的に理解するために，心理学や神経科学などの分野で様々なアプローチが行われてきた．例えば，心理学的なアプローチでは，スリルのある職業と成績などの関連からスリル欲求の傾向を分析した Farley の “Type T” がある [2]．また神経科学的なアプローチでは，ストレスなどの負の信号を予測し，ドーパミンニューロンを抑制する “手綱核” が発見されており，スリルのような本能的には不快である信号との関連が示唆された [3]．

しかしながら，スリルという概念を明確に定義している研究は少なく，これまでの研究の多くは単に人の行動の傾向分析や神経伝達メカニズムの解析が行われているに過ぎなかった．それゆえに，スリルを楽しむということを包括的に理解することができなかった．

競争の中でスリルを求めるという認知過程は，生得的には危険であることを察知しながらあえてその危険を冒すことを楽しむという矛盾を孕んでいる．この状態価値をモデル化するためには，単一の報酬関数に基づく合理的な適応手法だけでは不十分で，複数の報酬要因とそれらの相互作用系を適切に設計する必要がある．そこで本研究では，2 つの相反する報酬系を用意し，それらの相互作用に基づいた状態価値を獲得することで，スリルの認知モデルを実現することを目的とする．スリルは (a) 生得的な状態価値と (b) 経験的な状態価値によって認知されると仮定し，この 2 つの状態価値をそれぞれ (A) 進化計算と (B) 強化学習によりモデル化する．本研究の成果は，スリルという非合理的なリスク行動を理解できるという点において人と同調できるエージェントの実現に寄与し得る．

2. スリルのモデル化

本能的には不快であるはずの状態が楽しむ対象として経験されている状況をスリルを楽しむ状態としたとき，その行動要因には次の生得的・経験的な状態価値が関与する（図 1）．

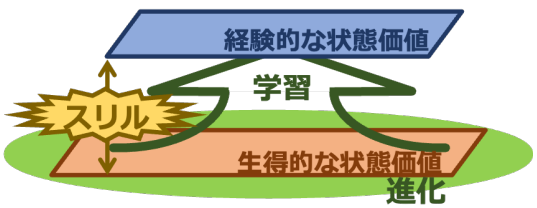

図 1　2 つの状態価値とその差分としてのスリル

生得的な状態価値 対象の状態は本能的に不快である
経験的な状態価値 対象の状態は快感として学習されている

本研究ではシミュレーション実験により，以上の生得的・経験的な 2 つの状態価値を進化・学習シミュレーションによってボトムアップに構築した．生得的には不快でありつつもそれを快感として経験するためには，状態価値の探索的な変化が必要である．またスリルを楽しむためには，実際に死んだり怪我を負わないようにする必要があり，そのためには “もしこれ以上危険な状態に近づいたら死んだり怪我を負ってしまう/しまっていたかもしれない” といった仮想現実を経験的に創り出すことが重要であると考えられる．Watson and Szathmáry (2016) は，進化と学習はパラメータのランダムな変化（突然変異や確率的探索）を適応的プロセスに適用しているという点で共通し，その一方で，仮想的な状態を予測するか否かという点で異なると主張した [4]．すなわち，学習は進化と異なり将来の状態を予測しながら適応し，進化は結果的に生存した個体集団によって適応される．これらを踏まえれば，突発的な変化が必要な生得的・経験的な状態価値を獲得するために進化と学習を用いることは妥当であり，将来の期待報酬を予測しながら適応する学習を経験的な状態価値の獲得に用いることが適していると考えられる．

Sutiono et al. (2014) は，遊びの楽しさは連続的に変化することを示した [5]．これを踏まえれば，スリルの楽しさを状態価値として表現するにあたり，状態価値が連続的に変化するものとして考えることが望ましい．そこで本研究では状態価値の数値表現に，空間内の全ての座標に対して連続的な状態価値を割り当てる影響マップ（Influence Map）[6] を用いた．影響マップは 2 次元空間を各座標における状態価値でマッピングしたデータを表し，影響度の変化はガウス分布とした．

モデル化の流れを図 2 に示す．モデル化の手順は (A) 進化計算と (B) 強化学習の 2 段階に分かれる．進化計算は集団探索により大域的に解空間を探索し，強化学習は個体探索により経験的に学習された状態価値を探索する [7, 8]．(A) まずは進化計算によって危険な状態を避ける生得的な状態価値を獲得する．(B) その後，A によって獲得した状態価値を初期値として強化学習を行い，危険な状態に近づく経験的な状態価値を獲得する．この A と B の状態価値の差分をスリルの度合いと考えることができる．本研究では，進化計算手法として遺伝的アルゴリズム（Genetic Algorithm, GA）[7]，強化学習手法として Q 学習 [9] を用いた．

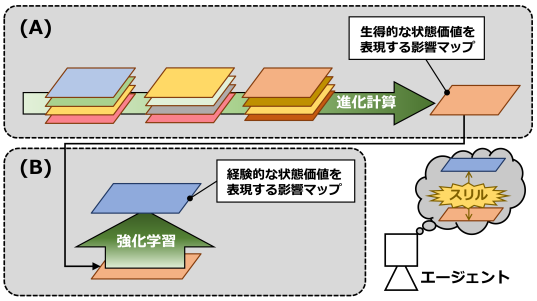

図 2　スリルのモデル化手順

3.　シミュレーション実験

3.1　シミュレーション環境

本研究では，2 次元のシミュレーション環境上で実験を行なった（図 3）．エージェントは 8 近傍への移動が可能で，左右どちらかのリンゴを目指して 50 ステップ移動する．50 ステップ経過した時点で課題は終了となり，エージェントは図 3 に示す初期位置に戻る．なお，環境はトーラス構造となっている．

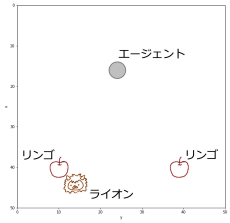

図 3　シミュレーション環境（エージェントが左右どちらかのリンゴを目指して移動する課題．エージェントは 8 近傍への移動が可能．環境はトーラス構造.）

3.2　進化計算

GA のハイパーパラメータを表 1 に示す．なお，GA の計算には Python の進化計算ライブラリである DEAP[10] を用いた．

表 1　進化計算（GA）のハイパーパラメータ

パラメータ名	値
個体数	8
世代数	10
交叉確率	0.9
突然変異率	0.2

図 3 に示したシミュレーション環境で，リンゴとライオンのそれぞれの状態価値を GA によって探索した．エージェントは，遺伝子の値（リンゴ/ライオンの状態価値）を混合ガウス分布における各平均として生成した影響マップの勾配を登るように移動し，リンゴに触れたら適応度を 1 に，ライオンに触れたら適応度を −1 に設定し，次世代選択・交叉を行なった．

GA の結果得られた状態価値とエージェントの軌道を図 4 に示す．図 4 より，リンゴには正の状態価値を，ライオンには負の状態価値を割り当てており，ライオンから離れたリンゴを取る行動を生成していることがわかる．また，得られた状態価値を表 2 に示す．表 2 より，リンゴの状態価値とライオンの状態価値の絶対値は非対称で，ライオンの状態価値の影響が大きいことがわかる．このように，ライオンの状態価値の影響をリンゴより大きくすることで左側のリンゴに近づかないようになるため，この結果は “危険を避ける” 合理的な状態価値を獲得したといえる．

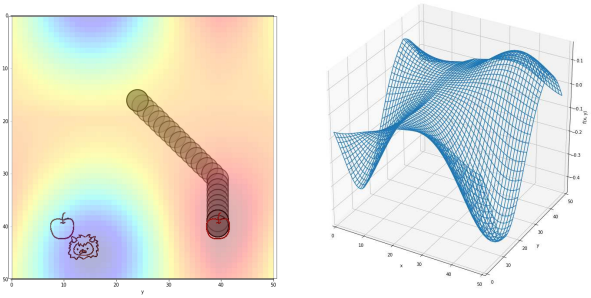

図 4　GA によって獲得された状態価値（左図は色が状態価値（赤が高く青が低い），右図は垂直軸が状態価値を表す）

表 2　GA によって獲得された状態価値

オブジェクト	値
ライオン	-0.601
リンゴ	0.167

3.3　強化学習

次に，図 4 の状態価値を状態価値関数の初期値として Q 学習を行った．Q 学習におけるハイパーパラメータを表 3 に示す．

表 3　強化学習（Q 学習）のハイパーパラメータ

パラメータ名	値
エピソード数	100,000
学習率 η	0.05
割引率 γ	0.85
探索確率 ϵ	0.2

報酬関数は以下の式 (1) として定義した．式 (1) において，x, y はエージェントの座標，$IM(x, y)$ は影響マップにおける座標 (x, y) の値である．また，w は “どの程度生得的な価値に逆向するか” を表すパラメータである．実験として，w に $(-1.0, -0.9, ..., 0.9, 1.0)$ の値をそれぞれ設定し，各パラメータごとに学習を行なった．

$$r = \begin{cases} -1 & (\text{ライオンに接触}) \\ -IM(x,y)w + 1 & (\text{リンゴに接触}) \\ -IM(x,y)w & (\text{上記以外}) \end{cases} \quad (1)$$

Q 学習の結果エージェントが得た累計報酬の推移を図 5 に示す．図 5 では，式 (1) における w の値ごとにプロットしている．この結果より，$w < 0$ の場合は 10,000 ステップ以内に学習が収束しているが，w が高くなると収束に時間がかかり，さらに $w = 1.0$ などは 40,000 ステップ付近で急激に獲得報酬が上昇していることがわかる．

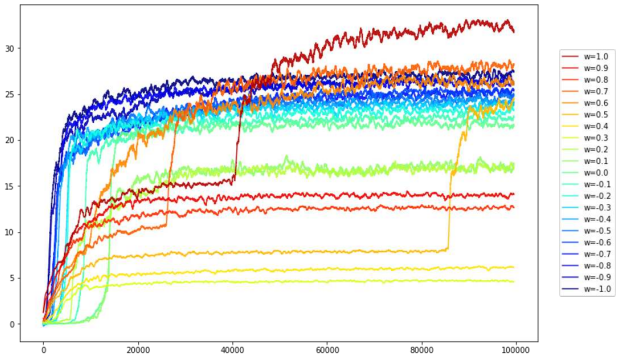

図 5　Q 学習の結果獲得した報酬の推移

Q 学習の結果得られた状態価値とエージェントの軌道を図 6 に示す．図 6 より，w の値に応じて以下の 4 パターンの状態価値の学習過程が確認された．

- 生得的な状態価値に従う $(w = -1.0 \sim 0.0)$
- 途中までライオンに向かう $(w = 0.1 \sim 0.2)$
- ライオンの手前で動き回る $(w = 0.3 \sim 0.4, 0.8 \sim 0.9)$
- ライオンに近づきつつリンゴを取る $(w = 0.5 \sim 0.7, 1.0)$

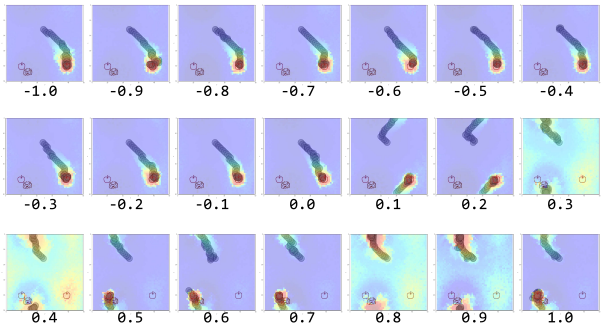

図 6　パラメータ w の変化による状態価値とエージェントの軌道

以上の 4 パターンの状態価値の獲得過程を図 7 に示す．図 7 では，4 つのどの条件においても左端の状態価値は図 4 に示した状態価値と同じであるが，学習エピソードが経過すると全く異なる状態価値のパターンに適応されていることがわかる．

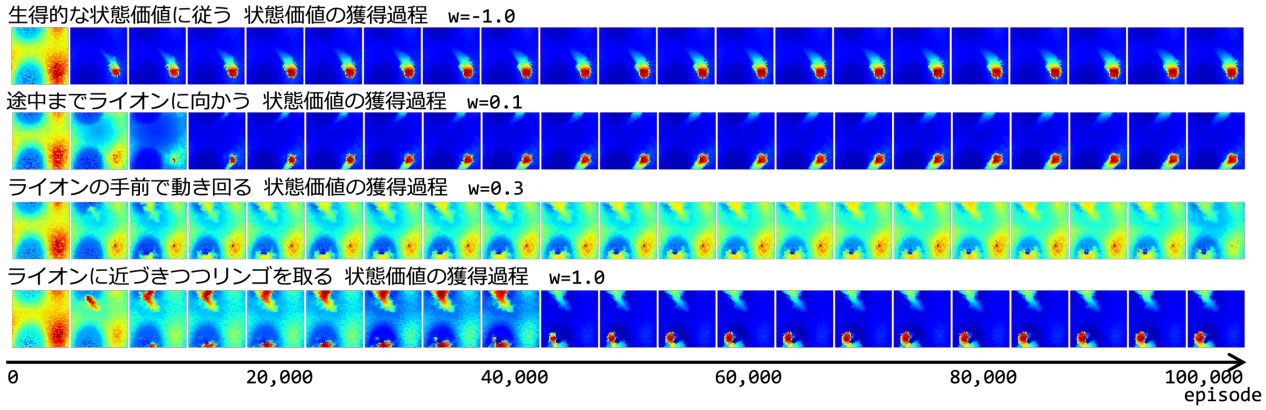

図 7　Q 学習による 4 パターンの状態価値の獲得過程（$w = -1.0, 0.1, 0.3, 1.0$）

4.　議論とまとめ

本研究では，進化計算と強化学習という適応的最適化手法を用いて状態価値を獲得させることでスリルをモデル化する手法を提案し，シミュレーション実験によってその有効性を確認した．実験の結果，進化計算によって獲得した状態価値は危険を避けるための合理的な行動を生成し，強化学習によって獲得した状態価値は重み w に応じて危険を冒す行動を生成した．このように，スリルという非合理的な認知活動は，相反する報酬関数とその重み w によってモデル化できることが示唆された．

Milano and Nolfi (2022) は，進化計算と強化学習の質的な違いのひとつは，解空間をどのように探索するか，だと主張した [11]．進化計算は少ない行動で到達可能な大域的な解空間を効率的に解く一方で，強化学習は多少複雑な行動で到達する特定の範囲の解空間の探索を行う．図 7 からわかるように，進化計算で獲得した状態価値は特定のタイミングで局所的に変化している．また，図 4 と図 6 のエージェントの軌道を比較しても，その複雑さは全く異なることがわかる．このような，進化計算と強化学習の適応過程と行動の質的な違いは，スリルの認知モデルを考えるうえで重要であると考えられる．

本研究では，避けるべき危険な状態（ライオン）は移動しない．一方で実世界において危険な状態は常にダイナミックに変化し続けている．このように状態価値が空間的に一定でない環境においては，提案モデルでは学習が収束しない可能性がある．そのため，今後は動的な環境で検証していく必要がある．

本研究の提案モデルでは，“危険を避ける” という状態価値をベースに学習することで，自身や他者の生死を尊重した行動生成が可能となり，人や他の生物と共生するエージェントの実現に向けた指針を示すことができると考えられる．また，これまでの単一な報酬関数だけでは表現できなかった状態表現を可能にするだけでなく，進化・学習といったプリミティブな生物性に立脚した人の行動原理の解明に寄与し得る．

謝辞

本研究は，科学技術融合振興財団（FOST）の助成を受けて行われた．ここに謝意を示す．

文献

[1] R. カイヨワ著, 多田道太郎, 塚崎幹夫訳, （1990）“遊びと人間”, 講談社学術文庫. （Roger Caillois, （1967）“Les Jeux et les Hommes”, Gallimard Education.）
[2] Farley, F., (1991) “The type-T personality”, Self-regulatory behavior and risk taking: Causes and consequences, pp. 371-382.
[3] Hikosaka, O., (2010) “The habenula: from stress evasion to value-based decision-making”, Nature reviews neuroscience, Vol. 11, No. 7, pp. 503-513.
[4] Watson, R. A., & Szathmáry, E., (2016) “How can evolution learn?”, Trends in Echology & Evolution, Vol. 31, Issue. 2, pp. 147-157.
[5] Sutiono, A. P., Purwarianti, A., & Iida, H., (2014) “A mathematical model of game refinement”, In Intelligent Technologies for Interactive Entertainment: 6th International Conference.
[6] Tozour, P., (2001) “Influence mapping”, Game programming gems, Vol. 2, pp. 287-297.
[7] Bäck, T., Fogel, D. B., & Michalewicz, Z., (2018) “Evolutionary computation 1: Basic algorithms and operators”, CRC press.
[8] Sutton, R. S., & Barto, A. G., (2018) “Reinforcement learning: An introduction”, MIT press.
[9] Watkins, C. J., & Dayan, P., (1992) “Q-learning”, Machine learning, Vol. 8, pp. 279-292.
[10] Fortin, F., De Rainville, F., Gardner M., Parizeau, Marc., & Gagné, C., (2012) “DEAP: Evolutionary Algorithms Made Easy”, Journal of Machine Learning Research, Vol. 13, pp. 2171-2175.
[11] Milano, N., & Nolfi, S., (2022) “Qualitative differences between evolutionary strategies and reinforcement learning methods for control of autonomous agents”, Evolutionary Intelligence, pp. 1-11.