進化計算と強化学習によるスリルのモデル化

A Cognitive Model of Thrill Using Evolutionary Computation and Reinforcement Learning

高田 亮介 [1*]　坂本 孝丈 [2]　竹内 勇剛 [2]
Ryosuke Takata[1]　Takafumi Sakamoto[2]　Yugo Takeuchi[2]

[1] 東京大学　[2] 静岡大学
[1] University of Tokyo　[2]Shizuoka University

Abstract: 人はしばしば生得的に危険を察知しつつもあえてその危険を冒し，スリルを楽しむという理知的には矛盾した非合理的ともいえる行動をとることがある．このような快感を伴うリスク行動は，単一の報酬関数に基づく合理性に従った進化計算や強化学習ではそれを適切に説明するモデルの構築は困難である．そこで本研究では，このような報酬要因が相反するリスク行動を求める主体の認知的状態のモデル化を目指す．シミュレーション実験では，ある行動に関して (a) 進化的にそれが危険であるという集団探索に基づいた価値と，(b) それが快感を生起させる対象として個体探索によって経験的に学習された価値の相互作用を通して，スリルを楽しむという非合理的な認知過程のモデル化の有効性を示唆する結果を得た．

1　はじめに

カイヨワは，人の遊びにおいて合理的に競争に勝つための行動は遊びの一面でしかなく，むしろ規則から外れて危険を冒すといった非合理的な行動こそが遊びの本質であると述べた [1]．例えば鬼ごっこ遊びで，あえて鬼に近づき挑発することで自らの危険を煽るリスク行動は，鬼ごっこの勝利という点では非合理的である．しかし人は危険であると理解している状態にあえて向かうことでスリルを味わうという遊びの楽しさを感じることができる．

これまで，遊びの中での人の行動プロセスを構成論的に理解するために強化学習を用いて行動生成モデルを作成する研究が行われてきた．その題材は，囲碁やチェス，将棋などのボードゲーム [2][3][4] や Atari，StarCraft といったコンピュータゲーム [4][5]，あるいは鬼ごっこ [6] など多岐にわたる．しかしながら，これらの研究ではゲームに勝利するという点で合理的な状態価値を獲得することを目的に強化学習を行っており，スリルを楽しむようなゲームの勝利に非合理的な状態価値の獲得過程を説明することはできていない．

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，これまで行われてきた強化学習のような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の適切なマネジメントを行う必要がある．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．

*連絡先：東京大学大学院総合文化研究科
〒 153-0041 東京都目黒区駒場 3-8-1
E-mail: takata@sacral.c.u-tokyo.ac.jp

2　スリルのモデル化手法

本来は不快であるはずの恐怖感が楽しむ対象として変化した状態がスリルを味わう状態としたとき，その行動要因には以下の生得的・経験的な状態価値が関与する（図 1）．

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 2 つの状態価値をボトムアップに構築した．図 1 に示したように，生得的な状態価値は進化によって獲得され，経験的な状態価値は学習によって獲得されると考えられる．生得的には不快でありつつもそれを快感として

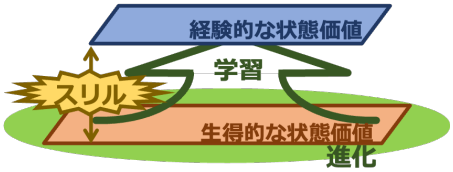


図 1: 2 つの状態価値とその差分としてのスリル

経験するためには，状態価値の突発的な変化が必要である．Watson and Szathmáry (2016) は，進化と学習はパラメータのランダムな変化（突然変異や確率的な探索）を適応的プロセスに適用しているという点で共通すると主張した [7]．これを踏まえれば，状態価値の適応に進化・学習は適していると考えられる．一方で，進化と学習は将来の状態を予測するか否かという点で異なるとも主張している．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 [7]．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創り出すことが重要であると考えられる．このように予測の中で危険に晒されるために，そこで，進化と学習のシミュレーションによって 2 つの状態価値をそれぞれモデル化した．状態価値は影響マップを用いて表現し，2 つの状態価値の差分をとることでスリルの認知を行うモデルを作成した．

モデル化の全体の流れを図 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．(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．

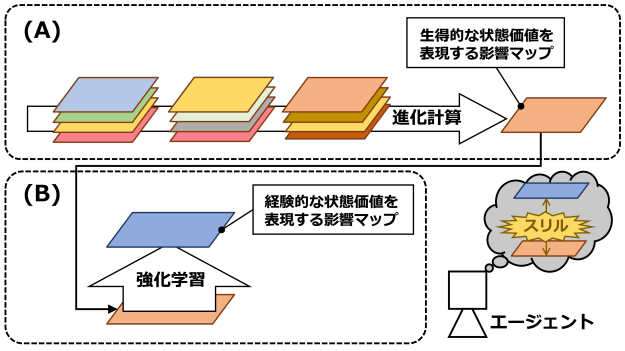


図 2: スリルのモデル化手順

2.1 影響マップ

Sutiono et al. (2014) は，ゲームの楽しさは連続的かつ加速度的に変化する数理モデルで表現可能であることを示した [8]．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，特定の状態だけでなくその周囲の状態価値も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影響マップ（Influence Map）[9] を用いた．影響マップは空間内の特定の位置に対して価値を割り当て，その周囲への影響度を計算して空間に適用する手法である．この手法により，連続的な状態価値の変化が表現可能になり，さらに視覚的な分析ができる．今回，影響度の変化はガウス分布とした．影響マップは，例えば図 3 のようになる．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であり，赤色は状態価値が高いことを表し，青色は状態価値が低いことを表す．

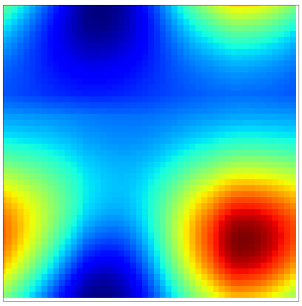


図 3: 影響マップの例

2.2 進化計算

ダーウィン進化のシミュレーション手法である進化計算は，集団探索により大域的に解空間を探索できる．本研究では，代表的かつ単純な進化計算手法である遺伝的アルゴリズム（Genetic Algorithm．以降，GA）を用いた [10]．

進化計算では生得的な状態価値の獲得を目指す．そのため，エージェントは環境に対する状態価値の一切を持たない状況から，環境とのインタラクションの帰結（例えば，ライオンに向かって歩いたらライオンに襲われて怪我を負った，など）をもとに，環境に対して状態価値を割り当てていく．この操作を行う個体集団を用意して進化アルゴリズムを適用することで，危険を避ける合理的な状態価値を集団探索することができる．

2.3 強化学習

大脳基底核の働きをモデル化した強化学習は，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ことができる [11]．本研究では，代表的かつ単純な強化学習手法である Q 学習を用いた [12]．Q 学習は，式 (1) に示すように Q 値を更新していく学習手法である．ここで，s_t は時点 t における状態，a_t は時点 t における行動，r_{t+1} は時点 $t+1$ における即時報酬，η は学習率，γ は割引率を表す．

$$Q(s_t, a_t) \leftarrow (1-\eta)Q(s_t, a_t) + \eta(r_{t+1} + \gamma max_{a_{t+1}} Q(s_{t+1}, a_{t+1})) \quad (1)$$

強化学習では経験的に危険に向かう状態価値の獲得を目指す．エージェントがスリルを味わうためには，エージェントは危険を冒しつつも最終的には危険を回避しなければならない．そのため，エージェントの学習が行われるタイミングで，前節で述べた進化計算によって獲得された危険を避ける状態価値を用いる．エージェントに報酬が与えられた際（例えば，食べ物にありついた，など）に，危険な状態であるほど報酬が高くなるように報酬関数を設計する．これにより，危険な状態でありつつも最終的には課題を達成する，というスリルを味わう条件を満たすことができる．

3 シミュレーション

3.1 環境

本研究では，実験として図 4 に示す 2 次元仮想環境を作成してシミュレーションを行った．図 4 には，エージェントとリンゴ，ライオンが存在する．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 ステップ移動する課題である．初期段階ではリンゴとライオンに対する状態価値は存在せず，リンゴに触れると報酬が与えられ，ライオンに触れるとペナルティが与えられるといった環境の規則のみ存在する．なお，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 4 に示す初期位置に戻る．

また，本実験で用いた計算環境を表 1 に示す．

表 1: 計算環境

名称	値
OS	Ubuntu 20.04.5 LTS
CPU	Intel Xeon 2.20GHz
GPU	Tesla T4
RAM	12GB

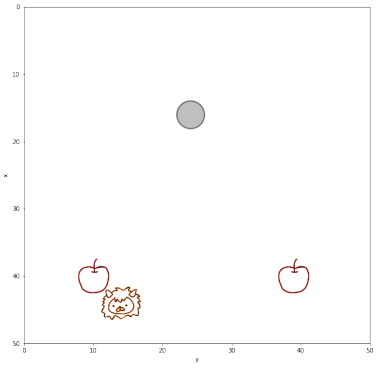


図 4: シミュレーション環境（黒い丸はエージェント．左右どちらかのリンゴを目指して移動する課題．エージェントは 8 近傍への移動が可能．）

3.2 進化計算実験

GA におけるハイパーパラメータを表 2 に示す．なお，今回は Python の進化計算ライブラリである DEAP[13] を用いた．

表 2: 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
個体の突然変異率	0.1
遺伝子の突然変異率	0.2
遺伝子の突然変異の平均	0.0
遺伝子の突然変異の標準偏差	0.5

図 4 に示した環境で，リンゴとライオンのそれぞれの状態価値を GA で探索した．探索する遺伝子長は 2 で，(1) ライオンの状態価値，(2) リンゴの状態価値である．解集団は，遺伝子から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・交叉を行った．

GA の結果得られた状態価値を図 5 に示す．図 5 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いることがわかる．また，得られた状態価値を表 3 に示す．表 3 に注目すると，リンゴの状態価値とライオンの状態価値は 0 に対称ではなく，ライオンの状態価値の影響が大きいことがわかる．ライオンの状態価値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，この結果は合理的であるといえる．

また，このときに生成されたエージェントの軌道を図 6 に示す．このエージェントの軌道からも，ライオ

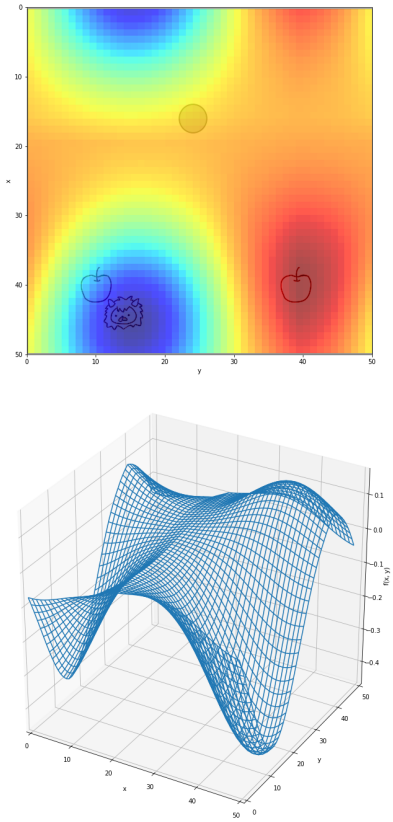


図 5: GA によって獲得された状態価値（上図は色が状態価値，下図は垂直軸が状態価値を表す）

表 3: 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	-0.6006432868757949
リンゴ	0.1667640789100624

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．以上より，進化的に獲得された状態価値は危険を避ける合理的な行動を生成することが示唆された．

3.3 強化学習実験

次に，図 5 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 4 に示す．

報酬関数は以下の式 (2) として定義した．式 (2) において，x, y はエージェントの座標，$IM(x, y)$ は影響マップの座標 (x, y) における値であり，$IsTouchLion(x, y)$ は座標 (x, y) がライオンに触れていれば 1，触れていなければ 0 を返す関数である．

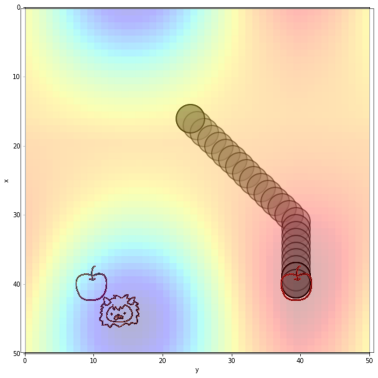


図 6: GA によって獲得されたエージェントの軌道

表 4: 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

$$r = \begin{cases} -IM(x, y) & (IsTouchLion(x, y) = 0) \\ -1 & (IsTouchLion(x, y) = 1) \end{cases} \tag{2}$$

Q 学習の結果エージェントが得た累計報酬の推移を図 7 に示す．図 7 より，およそ 20,000 ステップで学習が収束していることがわかる．

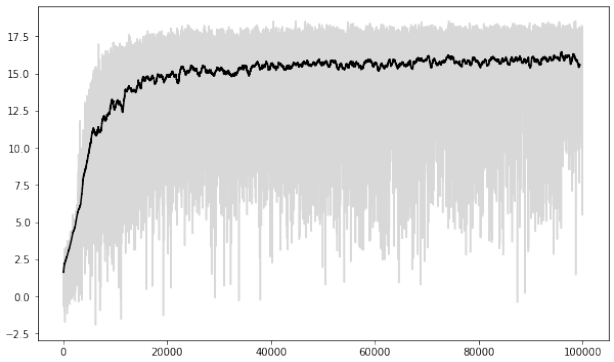
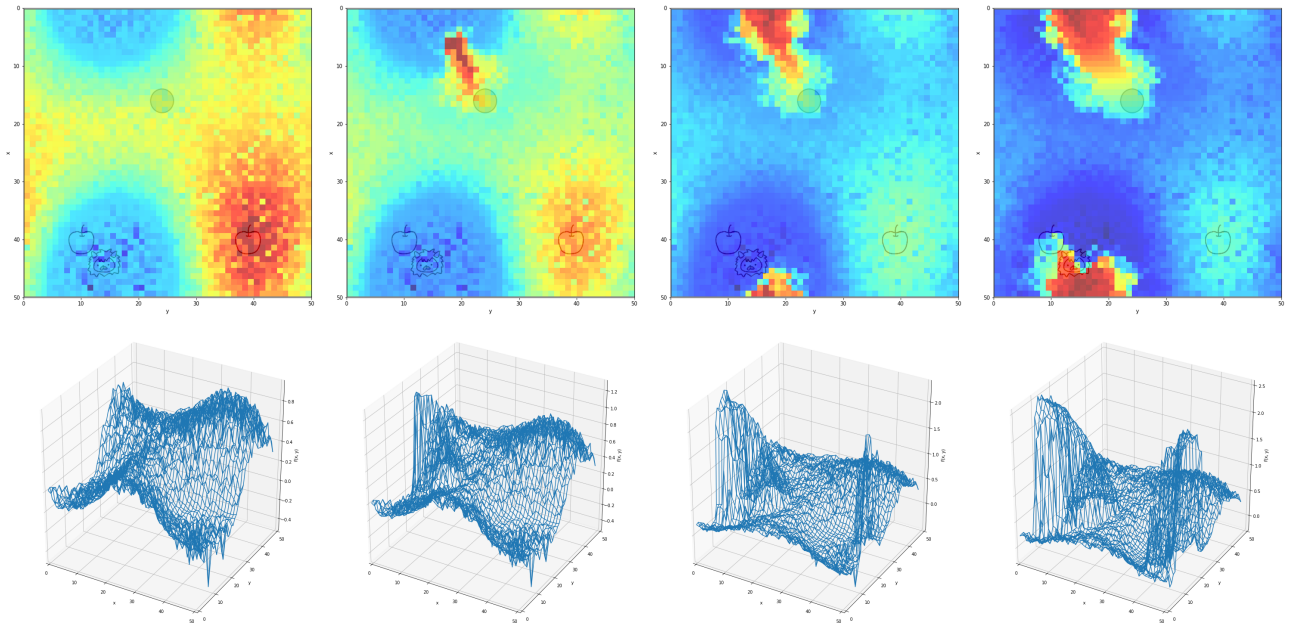


図 7: Q 学習の結果獲得した報酬の推移

Q 学習の結果得られた状態価値を図 8 に示す．図 8 では，左から 0 エピソード，4,000 エピソード，10,000 エピソード，100,000 エピソード時点での状態価値を表す．図 8 より，学習初期段階の状態価値は GA で獲得した図 5 に近いが，エピソードが進むにつれてそれまで状態価値の低かった左のリンゴ付近の状態価値が高くなっていることがわかる．100,000 エピソードではラ



(a) 0 エピソード　(b) 4,000 エピソード　(c) 10,000 エピソード　(d) 100,000 エピソード

図 8: Q 学習によって獲得された状態価値の遷移（上図は色が状態価値，下図は垂直軸が状態価値を表す）

イオンの近くまで状態価値が高くなり，対して右側のリンゴに対する状態価値は低くなっている．また，最終エピソードでのエージェントの軌道を図 9 に示す．図 9 より，GA によって低い状態価値が割り当てられた左側のリンゴに向かう行動が生成されていることがわかる．この結果は，進化的に獲得した状態価値から生成される行動とは対照的であり，生得的には危険であると察知する状態に向かう学習過程を実現できたことが示唆された．

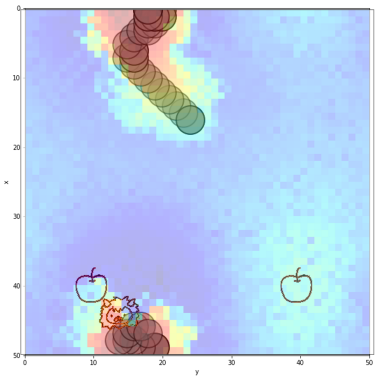


図 9: Q 学習によって獲得されたエージェントの軌道

4　議論

4.1　進化と学習による状態価値

シミュレーション実験より，進化計算では危険を避けるための合理的な状態価値を獲得し，強化学習ではその危険を避ける状態価値を初期値としながらも危険を冒す状態価値を獲得した．図 8 から見てとれるように，強化学習の初期段階（図 8(a)(b)）では進化計算によって獲得した状態価値が全体的なパターンとして見られるが，学習が進むにつれて状態価値が局所的に変化している（図 8(c)(d)）．最終的に獲得した状態価値（図 8(d)）では，左のリンゴ付近に高い状態価値があるが，ライオンとの接触を避けるような状態価値の形状となっている．このように，強化学習による状態価値は細かい変化が行われていることがわかった．

Milano and Nolfi (2022) は，進化計算と強化学習の質的な違いについてまとめている [14]．そこで述べられていることのひとつに，解空間をどのように探索するか，という点がある．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．この違いは今回のスリルを獲得するための 2 つの状態価値の探索に対応していると考えられる．すなわち，危険を避けるという大域的な行動を生成する状態価値の獲得に進化計算を用いて，危険であることを察知しながらもその危険を冒すという単調でない状態価値の獲得に強化学習を用いたことは，上記の知見と合致する．

4.2 スリルの学習過程

Doya (2002) は，強化学習における学習率や割引率といったハイパーパラメータが脳の神経伝達物質と対応していることを示唆した [15]．本実験では表 4 に示したハイパーパラメータのみでシミュレーションを行ったが，様々なハイパーパラメータでどのような行動変容が生じるかを分析することで，スリルを味わう際の脳活動原理を構成論的に説明できる可能性がある．

5 おわりに

本研究では，進化と学習により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，進化によって獲得した状態価値は危険を避ける行動を生成し，学習によって獲得した状態価値は危険に向かう行動を生成した．重要な点は，学習の状態価値の初期値は進化によって獲得した状態価値を用いていることである．環境に対して全く状態価値がない状況からの学習とは異なり，生得的に獲得した状態価値のもとで学習することで非合理的な状態価値の表現が可能であることが示唆された．

スリルを楽しむことの背後には，それによって死んだり怪我をしないという信念があるといえる．したがって，信念の強さを変数として学習するモデルにすることが望ましい．今回のモデルでは排反する複数の報酬の相互作用によりリスク行動を学習させることができたが，信念の表現までは至っていないため，今後の課題としたい．また，避けるべき危険な状態（ライオン）が移動しない点も問題として挙げられる．実世界において，危険な状態はダイナミックに変化し続けている．そのように状態価値が一定でない環境においては学習が収束しない可能性があり，今後検証していく必要がある．

エージェントがスリルのような非合理的なリスク行動を人と同じように理解し同調できるようになれば，非合理的な行動から生じる遊びの楽しさを人に提供し共有することが可能となる．また，提案モデルでは生物に共通する “危険を避ける” という状態価値をベースに学習していることにより，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．このように，本研究の提案モデルは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．

参考文献

[1] R. カイヨワ著, 多田道太郎, 塚崎幹夫訳: 遊びと人間, 講談社学術文庫 (1990) (Roger Caillois: Les Jeux et les Hommes, *Gallimard Education*, 1967)

[2] Silver, D., Huang, A., Maddison, C. J., Guez, A., Sifre, L., Van Den Driessche, G., Schrittwieser, J., Antonoglou, I., Panneershelvam, V., Lanctot, M., Dieleman, S., Grewe, D., Nham, J., Kalchbrenner, N., Sutskever, I., Lillicrap, T., Leach, M., Kavukcuoglu, K., Graepel, T., & Hassabis, D.: Mastering the game of Go with deep neural networks and tree search, *Nature*, Vol. 529, No. 7587, pp. 484-489 (2016)

[3] Silver, D., Hubert, T., Schrittwieser, J., Antonoglou, I., Lai, M., Guez, A., Lanctot, M., Sifre, L., Kumaran, D., Graepel, T., Lillicrap, T., Simonyan, K., & Hassabis, D.: A general reinforcement learning algorithm that masters chess, shogi, and Go through self-play, *Science*, Vol. 362, No. 6419, pp. 1140-1144 (2018)

[4] Schrittwieser, J., Antonoglou, I., Hubert, T., Simonyan, K., Sifre, L., Schmitt, S., Guez, A., Lockhart, E., Hassabis, D., Graepel, T., Lillicrap, T., & Silver, D.: Mastering atari, go, chess and shogi by planning with a learned model, *Nature*, Vol. 588, No. 7839, pp. 604-609 (2020)

[5] Vinyals, O., Babuschkin, I., Czarnecki, W. M., Mathieu, M., Dudzik, A., Chung, J., H. Choi, D., Powell, R., Ewalds, T., Georgiev, P., Oh, J., Horgan, D., Kroiss, M., Danihelka, I., Huang, A., Sifre, L., Cai, T., P. Agapiou, J., Jaderberg, M., S. Vezhnevets, A., Leblond, R., Pohlen, T., Dalibard, V., Budden, D., Sulsky, Y., Molloy, J., L. Paine, T., Gulcehre, C., Wang, Z., Pfaff, T., Wu, Y., Ring, R., Yogatama, D., Wünsch, D., McKinney, K., Smith, O., Schaul, T., Lillicrap, T., Kavukcuoglu, K., Hassabis, D., Apps C., & Silver, D.: Grandmaster level in StarCraft II using multi-agent reinforcement learning, *Nature*, Vol. 575, No. 7782, pp. 350-354 (2019)

[6] Baker, B., Kanitscheider, I., Markov, T., Wu, Y., Powell, G., McGrew, B., & Mordatch, I.: Emergent tool use from multi-agent autocurricula, *In 8th International Conference on Learning Representations* (2020)

[7] Watson, R. A., & Szathmáry, E.: How can evolution learn?, *Trends in Echology & Evolution*, Vol. 31, Issue. 2, pp. 147-157 (2016)

[8] Sutiono, A. P., Purwarianti, A., & Iida, H.: A mathematical model of game refinement, *In Intelligent Technologies for Interactive Entertainment: 6th International Conference* (2014)

[9] Tozour, P.: Influence mapping, *Game programming gems*, Vol. 2, pp. 287-297 (2001)

[10] Bäck, T., Fogel, D. B., & Michalewicz, Z.: Evolutionary computation 1: Basic algorithms and operators. *CRC press* (2018)

[11] Sutton, R. S., & Barto, A. G.: Reinforcement learning: An introduction, *MIT press* (2018)

[12] Watkins, C. J., & Dayan, P.: Q-learning, *Machine learning*, Vol. 8, pp. 279-292 (1992)

[13] Fortin, F., De Rainville, F., Gardner M., Parizeau, Marc., & Gagné, C.: DEAP: Evolutionary Algorithms Made Easy, *Journal of Machine Learning Research*, Vol. 13, pp. 2171-2175 (2012)

[14] Milano, N., & Nolfi, S.: Qualitative differences between evolutionary strategies and reinforcement learning methods for control of autonomous agents, *Evolutionary Intelligence*, pp. 1-11 (2022)

[15] Doya, K.: Metalearning and neuromodulation, *Neural networks*, Vol. 15, No. 4-6, pp. 495-506 (2002)