満足化と記録共有による対抗模倣の強化学習的モデリング
Reinforcement Learning Modeling of Emulation by Satisficing and Record Sharing

其田 憲明 †, 高橋 達二 ‡
Noriaki Sonota, Tatsuji Takahashi
† 東京電機大学大学院, ‡ 東京電機大学理工学部
Graduate School of Tokyo Denki University. School of Science and Engineering, Tokyo Denki University
tatsujit@mail.dendai.ac.jp

概要

本論文では, Risk-sensitive Satisficing (RS) モデルと RS モデルを用いた満足化と記録共有による対抗模倣のモデリングについて, goal-setting theory との関連性について検証することで, RS モデルと人間の学習傾向の関連性を把握することを目的とする. そして, ノイズの 1 種類として観測報酬に関してエージェント間で確率的な揺らぎが発生するバンディットタスクを用いることで, 確率的なノイズと疎な間隔での情報共有について, それらの要素がどのような影響を及ぼすか検証した結果, RS モデルと RS モデルによる満足化と記録共有の対抗模倣のモデリングには goal-setting theory との共通する点が見られた.

キーワード : satisficing, emulation, social learning, goal-setting theory

1. はじめに

本研究では人間の社会的学習の側面を機械学習の 1 分野である強化学習を用いることで再現し, その有用性と情報共有に関するノイズと疎な間隔での共有に関して goal-setting theory の観点から考察を行うことで人間の社会的な学習の解析を目的とする.

人間の社会的な学習の側面として, 他者の成功情報を用いる模倣学習が代表的である. 模倣学習の種類には他者の行動をそのまま模倣する imitation learning のほかに, 他者の結果を再現する emulation learning[1] (対抗模倣) が存在しており, その代表的な例として, 1 マイルを 4 分で走るという壁を越える選手が現れると他の選手もその壁を超え始める, といった現象が挙げられ, スポーツのような競争の他に企業や複数国間の技術競争などの様々な場面でこのような現象が見られる [2]. 機械学習における模倣学習は逆強化学習や教師あり学習を用いた imitation learning が主流となっているが, これらの学習方法は他者の行動系列データを多数必要とするため, 場合によっては情報の収集するのが困難となる場合がある. 一方で満足化 [3] を機械学習に反映した, Risk-sensitive Satisficing (RS) モデル [4] による社会的学習では他者の成功情報のみを用いることで, 必要とする情報量が状態数に比例することなく学習を行うことが可能であると示されている [5]. そのため, RS モデルによる社会的学習は emulation learning の性質を反映できていると考えられる [6].

他方で goal-setting theory では, 人間は取り組んでいる問題に対して適切な目標を与えられるとそのパフォーマンスをより向上させることが可能であると述べられている [7][8]. また, この理論はキーコンセプトとして “What is the minimum score you would be satisfied with ?” としており, このように目標を設定することが最良であると経験則として得られている. よって, ここにおける目標というものが満足化における基準と近しい性質を持つと捉えることが可能であると考えられる. この goal-setting theory では目標を持つことがより効果的なものとなるために, 適切な目標を与えることが重要であると述べられている. そのためには, 個人の能力に見合ったものにする, フィードバックやコミットメントを行うことによって, 難易度は維持しつつ達成が不可能はないものにする工夫が必要である.

この goal-setting theory と RS モデルを用いた社会的学習の関係として, 他者の成功情報というものは同じ設計のエージェントによるものであることから達成可能性は十分に高いこと, また, エージェントグループ内部で自律的に基準値を向上させるという傾向は目標をグループ内で適切な難易度へと近づけようとしていると解釈が可能である. しかし, 他者の成績を目標として設定する際に, 共有された情報には不必要な情報や欠損などが含まれていることがあり, そのことが目標の達成や情報共有による対抗模倣が困難にするということを考慮するべきである. 人間であれば認識や表現能力の限界から, 機械は通信過程におけるノイズ

や通信路の制限から情報の不要な情報や欠損が発生する [9] と考えられ, これらによって共有された目標が達成不可能なものになってしまうことは避けるべき問題である. しかし RS による対抗模倣のモデルがこれらの現象に対処可能であるかについて検証されてはいなかった.

従って本論文では, RS モデルと RS モデルを用いた満足化と記録共有による対抗模倣のモデリングについて, goal-setting theory に基づいて考察すること, ノイズとして観測報酬に対してエージェント間で確率的な揺らぎが発生しうるバンディットタスクを用いて確率的なノイズと疎な間隔による情報共有の影響を検証することで, RS モデルと RS モデルによる対抗模倣のモデリングと goal-setting theory との共通項を発見し, 人間の社会的学習の性質を解析することを目的とする.

2. Risk-sentive Satisficing

強化学習で一般的に用いられる ϵ-greedy では, エージェントの意思決定が確率パラメータである ϵ によって探索と活用の割合が管理される. それに対して人間は満足化という意思決定により探索と活用を動的に切り替えていると考えられる. その人間の満足化を強化学習に応用したのが Risk-sensitive Satisficing (RS) である.

2.1 Risk-sensitive Satisficing

RS は, i 番目の行動の試行回数 $n(a_i)$ とその行動に対する報酬平均 $E(a_i)$, そして報酬平均に対する基準値 $\aleph$ から, 式 1 によって RS 価値関数が定義される.

$$RS(a_i) = n(a_i)(E(a_i) - \aleph) \tag{1}$$

RS はこの RS 価値関数を最大化する行動 a_i を選択する.

基準値 $\aleph$ に加え, 試行量 n を用いることによって, 基準を満たしていない非満足状態においては楽観的探索を, 基準を満たしている満足状態においては悲観的活用を行うことが可能である.

2.2 記録共有による対抗模倣

エージェント N 体からなるグループの i 番目のエージェントの最大行動報酬平均を E_i^{best} とした時, 式 2 によって グループ内で自律的に基準値 $\aleph_t$ を更新する.

$$\aleph_t \leftarrow \max_i E_i^{best} \tag{2}$$

3. goal-setting theory

goal-setting theory では, ある仕事をこなしている人間に目標を与えることで 仕事のパフォーマンスが向上するという報告を始めとした, 人間と目標の関係性について述べられた理論である. ここで語られる目標は例えば林業であれば切り倒すべき木の本数といったものであり, このような目標設定はスポーツや心理療法, 創造性など幅広い分野に適応することが可能である.

goal-setting theory では目標のメディエータとして, 選択や注意力, 努力, 持続性, 適切な戦略の入手が挙げられており, 人は目標を通じてそれらの要素を向上させることで結果としてパフォーマンスを向上させるとされている. 次にモデレータとして次の要素を挙げている, 1 つ目に現状に対するフィードバックを行うこと, 2 つ目にコミットメントを行うこと, 3 つ目に個人の能力に見合った目標を設定すること, 最後に仕事を行う人へ快適な環境や支援を提供することである. これらの要素が目標とパフォーマンスの関係性を強くしているとされている.

目標によってパフォーマンスが向上する一方で, 達成不可能な目標は短期的にはより探索を促すことが可能であるが長期的にはパフォーマンスが低迷すること, 一般的な目標は必ずしも個人に適した目標とはならないため, パフォーマンス向上は見込めない場合も存在する.

4. 実験: K 本腕ベルヌーイバンディット

K 本腕ベルヌーイバンディット問題とは報酬が確率的に 0 もしくは 1 を与えられる K 個のバンディットが存在する問題である. この実験タスクでは報酬が確率的に定まるため, 有限回の試行において各エージェントごとに自身の観測情報に基づいて推定された各腕の報酬平均は異なる. 本実験ではこの確率的な報酬による各エージェントの報酬平均間に生じる揺らぎをノイズの一種とみなし実験を行う.

実験 1 では単体エージェントに対して数通りの目標を与えることで希求水準とパフォーマンスの関係性を観測する. 実験 2 では複数の RS エージェントを用いた情報共有による対抗模倣のモデリングにおいて, バンディット問題による観測情報へのノイズがどのような影響を与えるかを観測する. 実験 3 では, 疎な情報共有が対抗模倣のモデリングにどのような影響を与えるかを調べる.

4.1 実験 1: 単体エージェントにおける希求水準とパフォーマンスの関係

報酬確率を 0.1, 0.2, 0.3, ..., 0.9 とした $K = 9$ 本腕バンディットとする. このとき RS エージェントに与える基準値 $\aleph$ をそれぞれ [0.15, 0.25, 0.35, ..., 0.85, 0.95] とした場合に, RS エージェントに与えられた目標が最終的なパフォーマンスにどのような影響を及ぼすか観測する. これらの目標のうち, $\aleph = 0.85$ は最適な腕と次点の腕の確率の平均に位置するため, 次の式で定められた最適基準値となる [4].

$$\aleph_{opt} = (P_{\text{first}} + P_{\text{second}})/2 \tag{3}$$

また, $\aleph = 0.95$ は存在しているバンディットの最高報酬確率よりも高い設定となるため, 目標の達成が不可能となる基準値設定である.

1 試行を 1 エピソードとし, 1,000 エピソードを 1,000 回行った平均を結果とする.

4.2 実験 2: 対抗模倣のモデリングと難易度の関係

バンディットの確率は最小値を 0.1, 最大値を 0.9, 腕の本数を K とする. この K の値を変更することで行動範囲の増減が RS モデルによる対抗模倣のモデリングにどのような影響を与えるかを観測する. K 個存在するバンディットの確率の中央値が 0.5 となるよう, 等間隔に確率を設定する. 例えば $K = 5$ の場合の各バンディットの報酬確率は, 0.1, 0.3, 0.5, 0.7, 0.9 のようになる.

RS エージェントについて, 行動の価値推定はそれぞれ各腕ごとに記録される試行回数と報酬獲得回数から計算される獲得期待値を用いる. 事前情報として最適基準値 (式 (3)) を与えられた RS エージェントとエージェント数を 2, 3, 4, 5, 6, 7 体で比較を行い, 最適基準値 $\aleph_{opt}$ は存在するバンディットの報酬確率から式 (3) によって定める. エージェント間の記録共有は 1 試行毎に式 (2) によって行う.

1 試行を 1 エピソードとし, 5,000 エピソードを 1,000 回繰り返した平均を結果とする.

4.3 実験 3: 疎な情報共有

実験 3 では情報共有の間隔が対抗模倣のモデリングにどのような影響を与えるかを観測する. バンディットの報酬確率の設定は実験 2 と同様とする. 情報共有は実験 2 と同様に式 (2) で行い, その間隔 (interval) を 1, 10, 100, 1000 試行毎と設定し, バンディットが K 個存在する場合の各間隔による成績を比較する.

1 試行を 1 エピソードとし, 5,000 エピソードを 10,000 回繰り返した平均を結果とする.

5. 結果

5.1 実験 1: 単体エージェントにおける希求水準とパフォーマンスの関係

各基準値に対する平均獲得報酬と累計期待損失を図 1 に示す. 与えられた希求水準となる目標が高くなるにつれて最終的に高い報酬を得ることに成功していることがわかる. しかし, 達成不可能な目標が与えられた場合は目標を満たす行動が発見できないため, 最終的な成績が下がっていることがわかる.

5.2 実験 2: 対抗模倣のモデリングと難易度の関係

平均獲得報酬, 累計損失期待値, 共有基準値の時間発展についてそれぞれ $K = 5, 25$ の場合の結果を図 2 にまとめる.

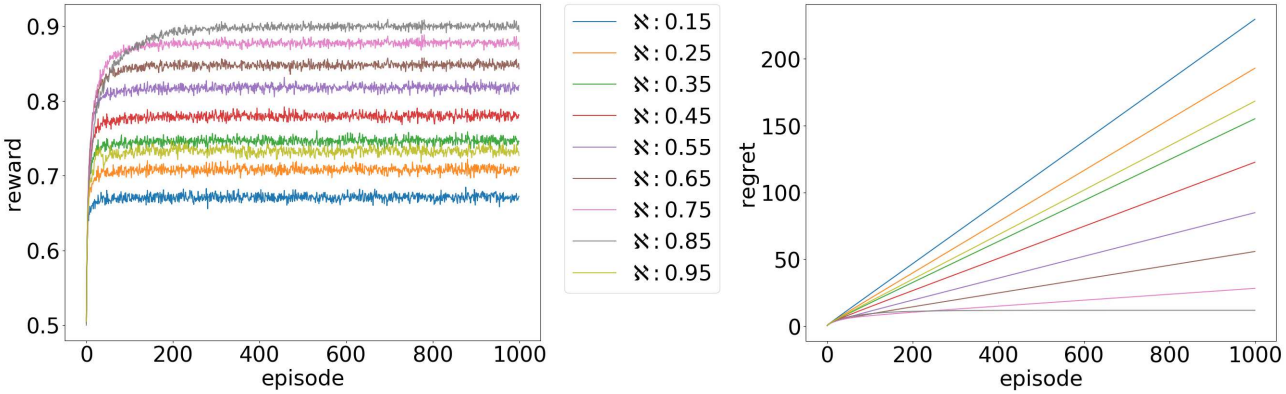
エージェント数が増えるごとに平均獲得報酬が向上しており, それに伴って累計損失期待値が減少していることがわかる. また, 共有基準値に関しても同様にエージェント数が増加するごとに高く設定することに成功しており, 今回の行った腕の本数の範囲ではエージェント数が 6 体程度で学習を行うことで最適基準を上回っていることがわかる.

5.3 実験 3: 疎な情報共有

情報共有間隔の差によるパフォーマンスの変化について図 3 に示す. 情報共有の間隔が疎であるほど基準値の上昇が遅くなり, 最終的な成績が低くなっていることがわかる.

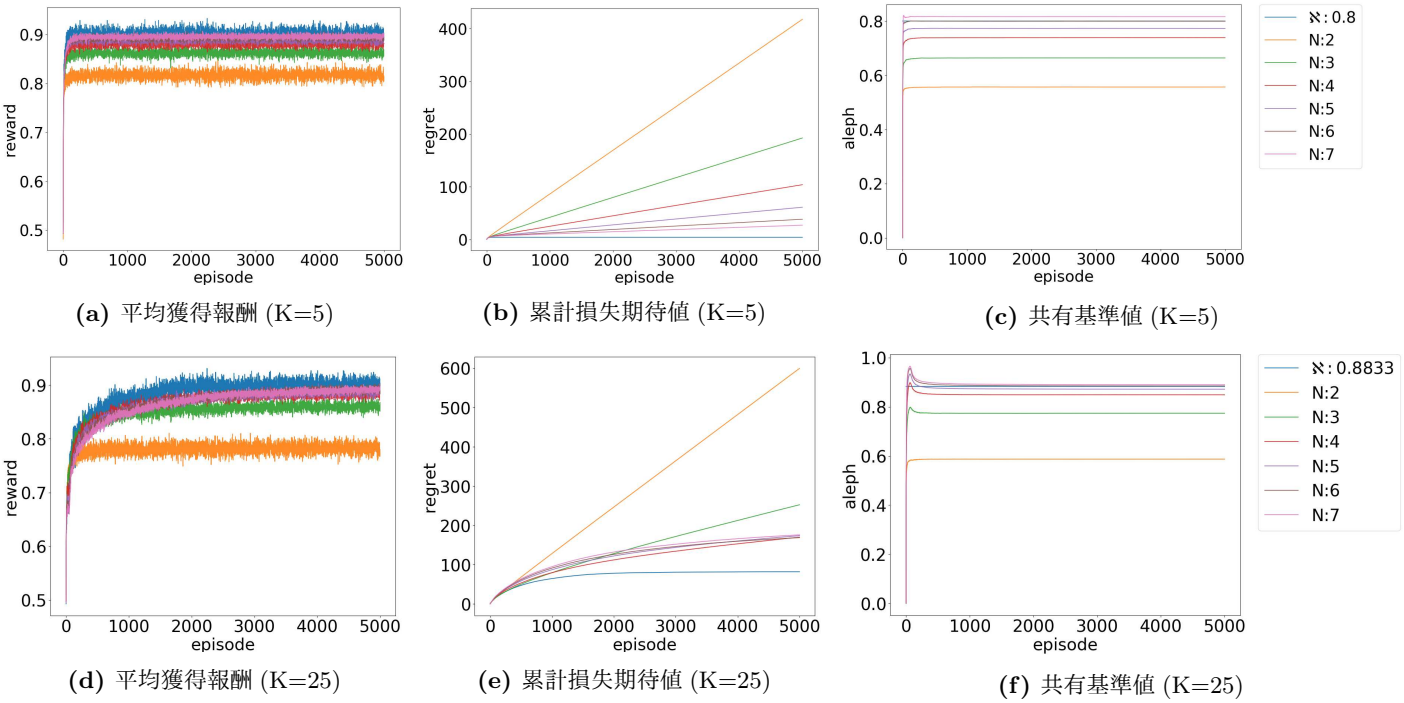
6. 考察

実験 1 では, 単体エージェントに対して目標となる基準値が高いほど良い成績を残すことに成功した. このことは goal-setting theory でも同様に高い目標であるほどパフォーマンスの向上に成功するという報告とも一致する. また, エージェントに与えられた目標が達成不可能なものである場合に成績が向上せずに落ち込んでしまうという結果も一致している. これらの結



(a) 平均獲得報酬 **(b)** 累計損失期待値

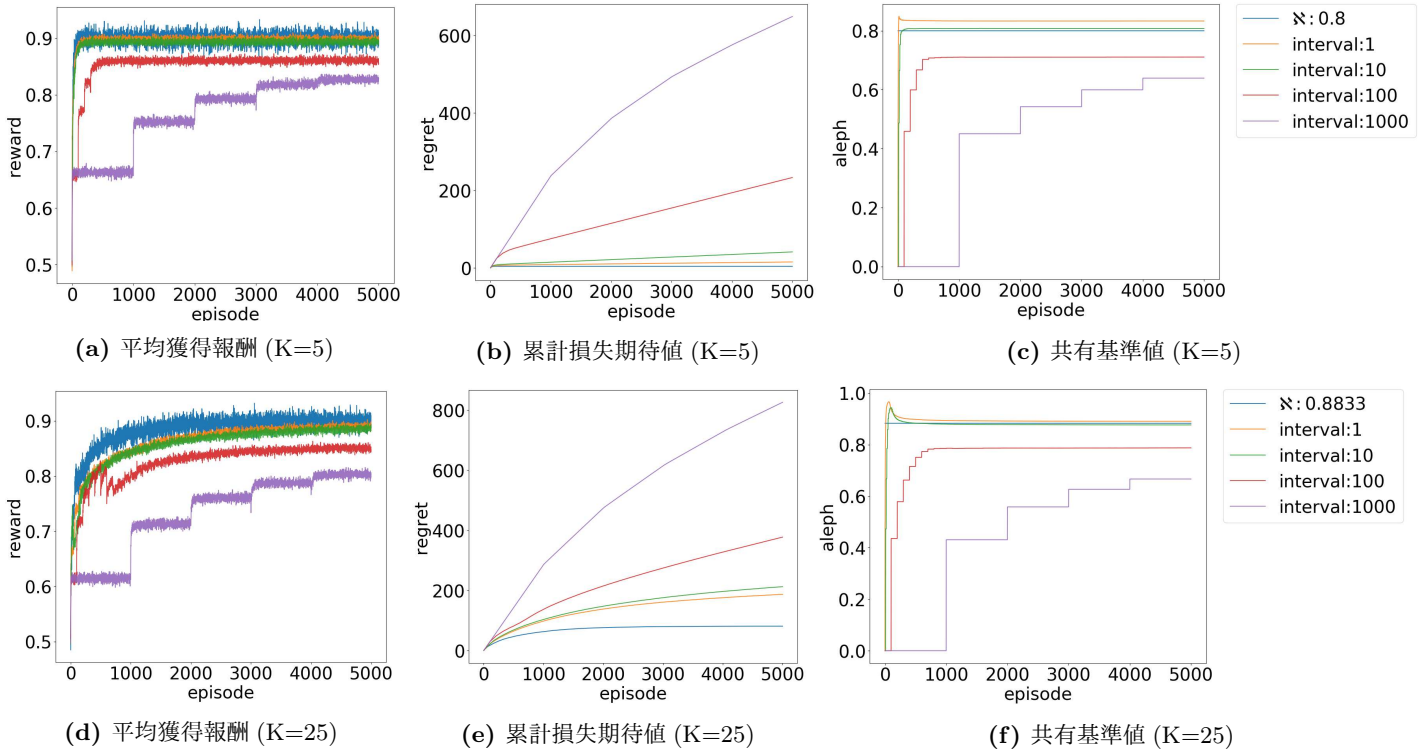
図 1: 実験 1: 単体エージェントにおける希求水準とパフォーマンスの関係



(a) 平均獲得報酬 (K=5) **(b)** 累計損失期待値 (K=5) **(c)** 共有基準値 (K=5)

(d) 平均獲得報酬 (K=25) **(e)** 累計損失期待値 (K=25) **(f)** 共有基準値 (K=25)

図 2: 実験 2: 対抗模倣のモデリングと難易度の関係



(a) 平均獲得報酬 (K=5) **(b)** 累計損失期待値 (K=5) **(c)** 共有基準値 (K=5)

(d) 平均獲得報酬 (K=25) **(e)** 累計損失期待値 (K=25) **(f)** 共有基準値 (K=25)

図 3: 実験 3: 疎な情報共有

果に対して, RS では与えられた目標が満たすことができない場合は常に探索を行い, 各行動の選び方が均衡的であるという性質があるため, 報酬確率が低い行動を定期的に選び続けることで累計損失期待値が線形的に上昇し続けると考えられる. goal-setting theory では人間が達成不可能な目標を与えられるとパフォーマンスが低迷する理由として, 成功体験という効用が得られないためであるとされており, RS エージェントも同様に基準を満たさない行動に対しては負の評価を行うため, 実現不可能な目標下では成功体験という基準を満たす行動による正の効用を得ることが不可能である. しかし, RS エージェントは成功体験を得られない間は基準を満たす行動を探索し続ける性質から, goal-setting theory における目標の高さと持続性や努力の関係には相関があるとは言い難いと考えられる.

実験 2 では, 観測情報にノイズが含まれる場合の探索空間とエージェント数の関係性について実験を行なった. エージェント数が多くなるほどグループの成績をより高くなるという結果が得られた. このことはエージェント数が増加する毎に環境に対するエージェントの探索が広い範囲で行われたことによるものだと考えられる. 同様にエージェント数を増やすことで累計損失期待値も減らすことに成功したが, 完全に上昇を抑えることに失敗している. これは各エージェントが観測している報酬が確率的なものであることから, 同じ選択肢から得られる情報がエージェント毎に異なること, そして共有される情報がグループ内で最も良い成績であるため, 今回用いたバンディットタスクでは常に確率的に良い報酬が得られたエージェントの情報を常に参照し続けることから, 他のエージェントは共有された成績を達成できることを保証されていないという性質がある. 従って実験 1 で達成不可能な目標を与えられたエージェントと同様の境遇にいるエージェントがグループ内に存在していると考えられる. この問題を回避するためには, 共有される情報に対して各エージェントが低めに補正するという方法が考えられるが, このことによってグループの成長が止まってしまう可能性を考慮する必要がある.

実験 3 では疎な情報共有とパフォーマンスの関係性について実験を行なった. そして, 情報共有の間隔が疎になるにつれて成績の向上が遅くなるという結果が得られた. goal-setting theory のモデレータとしてエージェントの目標はコミットメントやフィードバックが存在し, これらを通してよりパフォーマンスの向上を円滑にすることが可能であるとされている. このことから情報共有がされない, つまりコミットメントが行われないことによって, 各エージェントのフィードバックが適切なものになっていないと考えることが

できる. しかし, [8] ではパフォーマンスが向上した人間は個人でより高い目標を設定していたという結果も存在するため, 情報共有の他にエージェントが各自で基準値を更新する仕組みを取り組むことでより成績がより向上すると考えられる.

7. おわりに

本研究では, RS モデルと RS による対抗模倣のモデリングや情報に含まれるノイズや疎な間隔での情報共有が与える影響を goal-setting theory の観点から検証, 考察した.

RS エージェントが与えられた希求水準に比例して最終的なパフォーマンスが向上させることに成功したが, 達成不可能な目標はパフォーマンスを低減させる. 観測情報にノイズが含まれる場合には一部のエージェントが目標を達成できずに探索を続けるため, グループでの累計損失期待値の上昇を完全に抑えることが困難である. 情報共有が密であるほど到達できる最終的なパフォーマンスが高い. という 3 つの結果が得られた.

しかし, RS モデルはその設計上, 目標が達成できないことによる持続性, 努力の喪失といった現象が現れないといった人間とは異なる性質を持つため, 今後の課題として, これらの差異によって結果にどのような違いが現れるかを検証することが挙げられる.

文献

[1] Whiten, A. et al (2009): “Emulation, imitation, over-imitation and the scope of culture for child and chimpanzee” *Philosophical Transactions of the poyal society B*, 364(1528), 2417-2428. 364(2009)
[2] 柄谷行人, (1985). “ブタに生れかわる話”, 批評とポスト・モダン, pp. 257–60.
[3] Simon, H.A., (1956): “Rational choice and the structure of the environment”, *Psychological Review*, 63(2), 129–138.
[4] 高橋達二, 甲野佑, 浦上大輔, (2016): “認知的満足化 - 限定合理性の強化学習における効用”, 人工知能学会論文誌, 31(6), 1–11.
[5] 其田憲明, 神谷匠, 甲野佑, 高橋達二, (2019): “大局基準値共有による社会的強化学習”, JSAI 2019, 3K3-J-2.
[6] 高橋 達二, (2019): “神でもなく人間でもなく——現在の人工知能に何が足りないのか” (特集：人工知能と哲学・歴史・社会), 大学出版, 第 119 号, 18–23.
[7] Locke, Edwin A., Latham, Gary P.(2002): “Building a Practically Useful Theory of Goal Setting and Task Motivation”, *American Psychologist*, 57(9), 705–717.
[8] Locke, E.A., Latham, G.P. (2019): “The development of goal setting theory: A half century retrospective.” *Motivation Science*, 5(2), 93-105.
[9] 山川 宏, (2018): “機械知能社会は技術的特異点を超えられるか”, 人工知能, 33, 6, 873–879.