様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

令和　５ 年　 ６ 月 ２１ 日現在

機関番号：５４６０１
研究種目：基盤研究(C)（一般）
研究期間：2020～2022
課題番号：２０Ｋ１１９４６

研究課題名（和文）多目的強化学習の学習結果全ての分布を可視化する報酬生起確率ベクトル空間の構築

研究課題名（英文）Reward occurence probability vector space that Visualizes the distribution of whole learning results of multi-objective reinforcement learning

研究代表者

山口　智浩（Yamaguchi, Tomohiro）

奈良工業高等専門学校・情報工学科・教授

研究者番号：００２４０８３８

交付決定額（研究期間全体）：（直接経費）　　3,300,000円

研究成果の概要（和文）：まず，全ての報酬獲得方策の収集・多目的最適方策決定の並列化と部分計算による高速化を実装した．状態数12，報酬数3の確率的MDP環境で，報酬獲得方策数は25.3万に対し，報酬生起確率ベクトル数は5430と約1/50に減少した．報酬数4の場合，報酬獲得方策全てに対応する生起確率ベクトル集合の算出までに要する実行時間を従来手法と比較した結果，既存手法(1590秒)と比べ並列化手法(8.8秒)は，1/180に高速化された．次に，報酬数n=3の場合について，多目的最適方策を最適化するための目的間の重みベクトルの範囲の決定をメッシュ法で実現し，「重みベクトルに対する最適方策の平均報酬の可視化を実現した．

研究成果の学術的意義や社会的意義
本研究の学術的意義は，従来手法では，平均報酬最大となる多目的最適方策の境界を解析的に解くのが，目的数3以上の場合に困難だったのに対し，本手法では，各重みベクトルに対して，式(1)を用いて各方策の平均報酬値を算出し，最大となる方策を決定するため，計算コストの許す限り，近似的な算出が可能な点である．しかも，多目的最適方策の決定過程において，多目的間の重要度を表す重みベクトルとは独立な，報酬生起確率ベクトルをまず算出し，次にそれを用いて多目的最適方策を最適化するための，目的間の重みベクトルの範囲の決定を，メッシュ法を用いて近似的に行うことで，目的数3以上の場合の算出を実現した点である．

研究成果の概要（英文）：First, we implemented parallelization of the collection of all reward acquisition policies and the determination of the multi-objective optimal policies, as well as speeding up the process by partial computation. In a stochastic MDP environment with 12 states and 3 rewards, the number of reward acquisition policies was 253,000, while the number of reward occurrence probability vectors was reduced to 5430, about 1/50. In the case of 4 rewards, the parallelized method (8.8 sec) was 1/180th faster than the existing method (1590 sec) in terms of the execution time required to calculate the set of occurrence probability vectors corresponding to all reward acquisition policies. Next, for the case of 3 rewards, we used the mesh method to determine the range of weight vectors among the objectives to optimize the multi-objective optimal policy, and visualized the average reward of the optima policy for the weight vectors.

研究分野：強化学習

キーワード：機械学習　多目的強化学習　報酬生起確率ベクトル　重みベクトル　部分計算　多目的最適方策　可視化　ベクトル空間

様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９（共通）

１．研究開始当初の背景

本研究で用いる強化学習法は，学習目標を環境中の報酬として設定し，学習者が試行錯誤を通して自律的に目標への行動系列を学習する手法で，ロボットやエージェントの行動学習として広く用いられている．しかしながら，既存研究の目的は，最適解やパレート最適解等なんらかの最適性基準を満たす解の獲得であり，それら以外の非最適解に注目し，研究目標とする研究は，我々の研究以外には申請者の知る限り行われていない．
また，近年，注目されている深層(強化)学習は，(a)大量データからの学習と(b)学習過程での内部表現の自動生成を特徴とするが，その反面，(c)学習結果の根拠や理由が説明されないため，人が納得や理解できない，(d)自動生成された内部表現や学習過程が理解困難，という弱点がある．これに対し本研究では空間上での学習結果の幾何的な可視化を通して，大量の学習結果を人が理解しやすい可視化機能に持たせることで，深層学習の２つの問題点の解決を試みる．

２．研究の目的

本研究の目的は，機械学習の出力に対する根拠や理由を説明できる機構を多目的強化学習で実現し，その有効性を検証することである．具体的には，これまでの申請者らの基盤研究(C)研究で得られた成果を発展させ，多目的強化学習における学習結果全ての分布を可視化・説明する空間を構築[Yamaguchi 2020]し，目的間の任意の重みに対応した多目的最適方策を自動選択する機構の実現を目指す．

３．研究の方法

本研究では，以下の５項目の研究課題に取り組む．
(項目 1) 全ての報酬獲得方策を収集する多目的強化学習の並列化と部分計算による高速化
(項目 2) 全ての報酬獲得方策の分布を可視化するための報酬生起確率ベクトル空間の設計
(項目 3) 報酬生起確率ベクトル空間での多目的最適方策集合に基づくモデルの可視化
(項目 4) 多目的最適方策を最適化するための，目的間の重みベクトルの範囲の決定
(項目 5) 目的間の任意の重みに対応した多目的最適方策を自動選択する機構の評価

４．研究成果

初年度は，まず(項目 2)「全ての報酬獲得方策の分布を可視化する報酬生起確率ベクトル空間の設計」について，報酬数 n=3,4 の場合を検討した．n 個の報酬 Ri(i=1,2,…n)を要素とする報酬ベクトル R に対し，任意の方策が獲得する報酬の生起確率 pi(i=1,2,3,…n)を要素とするベクトルを報酬生起確率ベクトル P としたときに，任意の方策は，n 次元の報酬生起確率ベクトル空間内の１点で表わされる．重みの区間に応じて平均報酬最大となる方策集合は，空間の凸包の各頂点となり，既存の多次元凸包算出法で計算できる．

次に，(項目 1)「全ての報酬獲得方策の収集・多目的最適方策決定の並列化と部分計算による高速化」[Yamaguchi 2022] を実装・評価した．まず既存手法で全体の処理時間のボトルネックだった報酬獲得方策全ての収集は，n 個の報酬 Ri それぞれを起点とする木探索で行う．報酬別の木探索は並列化できるため，マルチプロセッシングによる CPU コア並列化を実装した．報酬数 n がコア数以下の場合，実行時間は最大 1/n となる．次に凸包算出の前処理として，収集した方策を生起確率ベクトルで多重ソートし，(異なる要素からなる)生起確率ベクトル集合を凸包算出前に部分計算した．状態数 12，報酬数 3 の場合，50 回の異なる確率的 MDP 環境において，平均の報酬獲得方策数 25.3 万(±8.3 万)に対し，平均の報酬生起確率ベクトル数は 5430(±5130)と約 1/50 に減少した．状態数 5～12，action 数 3，報酬数 n=3,4 の確率的 MDP 環境で報酬獲得方策全てに対応する生起確率ベクトル集合の算出までに要する実行時間を**図１**に示す．**図１**は，報酬数 n=3(実線)，n=4(点線)それぞれの場合について，報酬獲得方策全ての算出時間(**赤線**)(a)(b),および生起確率ベクトル集合の算出時間(**青線**)(c)(d)の 50 回の平均値である．従来手法 [Yamaguchi 2020] と比較した結果，状態数 12，報酬数 4 の場合，既存手法(1590 秒)と比べ並列化手法(8.8 秒)は，1/180 に高速化された．実行時間の概算は，コア並列化で最大 1/3～1/4，生起確率ベクトル集合の部分計算で約 1/50 だったので，両者を合わせると最大 150～200 倍の高速化が見込まれるため，実測値(1/180)は妥当である．

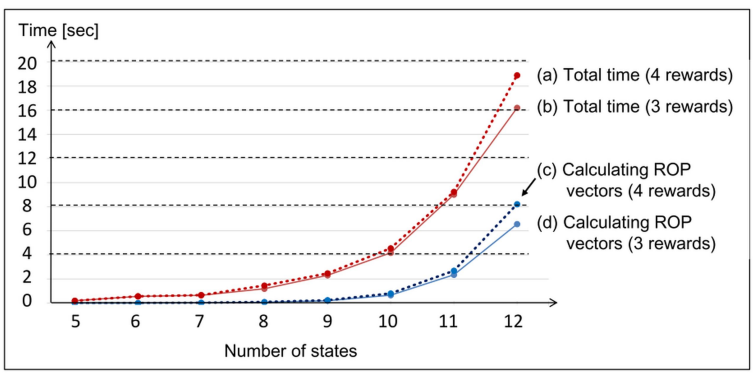

図１報酬獲得方策全て(赤線)に対応する生起確率ベクトル集合の算出(青線)に要する実行時間

２年目は，報酬数 n=3 の場合について，全ての報酬獲得方策の分布を可視化する報酬生起確率ベクトル空間における，(項目 4)「多目的最適方策を最適化するための目的間の重みベクトルの範囲の決定」および(項目 3)「報酬生起確率ベクトル空間での多目的最適方策集合に基づくモデルの可視化」について検討した．

まず，(項目 4)では，重み和＝1 となる制約を仮定すると，次元数を 1 減らすことができる．これを用いて既存手法では，報酬数 n=2 の場合には，2 次元重みベクトル空間における重みベクトルが x 軸となす角度 α を用いて，各最適方策の重みベクトルの境界が算出できることが知られていたが，報酬数 n=3 以上の場合には，解析的に解くことが困難であった．一方，本手法では，各方策の平均報酬が方策の報酬生起確率ベクトルと重みベクトルとの内積で算出されるため，方策ごとに再学習を要する既存手法と比べ，計算コストが小さい．これを利用して，最適重み区間を近似的に推定する手法を検討した．具体的には，メッシュ法を用いて重みベクトル空間を等間隔に分割し，メッシュ各点での各多目的最適方策の平均報酬を算出し，平均報酬最大となる方策を決定することで，各方策が最適となる重みベクトルの範囲を算出，可視化する方法を実装・評価した．**図 2** に報酬数 n=3，大きさ＝1 の，可視化された 3 次元の重みベクトルの分布を示す．図 2 の各軸は，3 つの報酬それぞれに対する重みを表す．

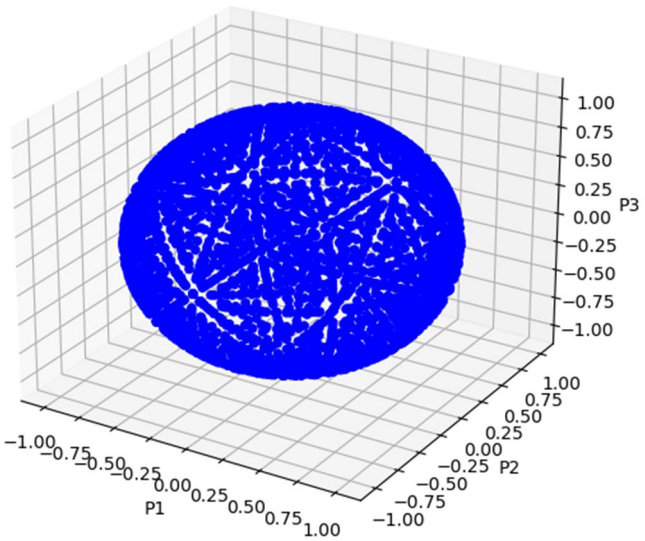

図 2　報酬数 n=3，大きさ＝1 の，可視化された 3 次元の重みベクトルの分布

では次に，(項目 3) モデルの可視化手法の分かり易さについて，上述のメッシュ法での重みベクトル空間の最適範囲の可視化手法を用いて検討した．まずメッシュ分割法を，重み値，重み比とで比較した結果，重み比に対して等分割する方が最適重み区間の境界付近の可視化が改善された．

次に，最適となる多目的方策の範囲を，(1)重み比の空間，(2)重みベクトル空間，(3) 重みベクトルに対する最適方策の平均報酬の可視化，の 3 手法について可視化した結果を述べる．なお，説明に用いた例は，報酬数 3 の場合での 6 種類の多目的最適方策を，赤，緑，青，黄，紫，水色の 6 色で可視化し，隣接する多目的最適方策の平均報酬が等しくなる境界を黒色で可視化している．

まず，**図 3** に(1)重み比の空間の可視化を示す．各軸は，メッシュ化に用いた重み比を表す．図 2 では，最適方策間の境界を表す黒点の分布が面状となり，境界値の可視化が明確であるが，各重みベクトルに対応する方策の平均報酬の大きさは可視化されていない．

次に**図 4** に(2)重みベクトル空間の可視化を示す．図 4 は，図 2 で示した重みベクトルについて，各重みベクトルに対応する多目的最適方策の色が割り当てられている．各軸は，メッシュ化に用いた重みベクトルの値を表す．図 3，図 4 は各重みベクトルに対応する多目的最適方策が色で可視化されている点が共通である．図 3 と比べ図 4 では，半径 1 の球面について，対応する多目的方策が一意に色で可視化されている点が特徴的である．但し，図 3 の場合と同様に，各重みベクトルに対応する方策の平均報酬の大きさは可視化されていない．

図 5 に(3)重みベクトルに対する最適方策の平均報酬$\rho_\pi(\vec{w})$ の可視化を示す．図 5 は，図 4 で可視化された色付けされた重みベクトルの大きさを，対応する最適方策の平均報酬値とした可視化である．ここで，$\rho_\pi(\vec{w})$ は，方策πが持つ報酬生起確率ベクトルと重みベクトル$\vec{w}$ との内積によって定義・算出される．定義式を式(1) に示す．以上 3 種類の可視化手法について，平均報酬の大きさと，方策が最適となる範囲の両方が可視化できる図 5 の手法(3)が最も分かり易いと判断した．

$$\boldsymbol{\rho_\pi(\vec{w})} = \sum_{i=0}^{d-1} \boldsymbol{w_i P_i} = \boldsymbol{\vec{w} \cdot \vec{P}_\pi} \qquad (1)$$

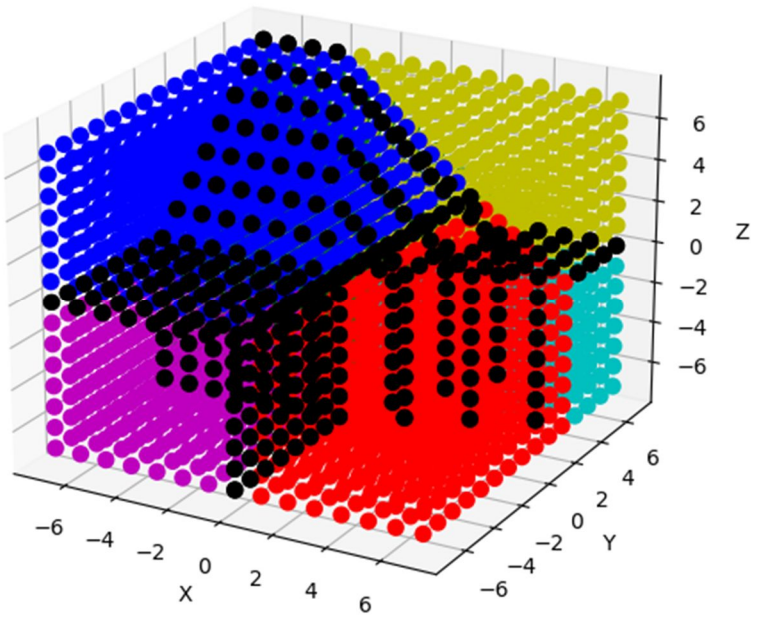

図 3　重み比の空間における最適方策分布の可視化

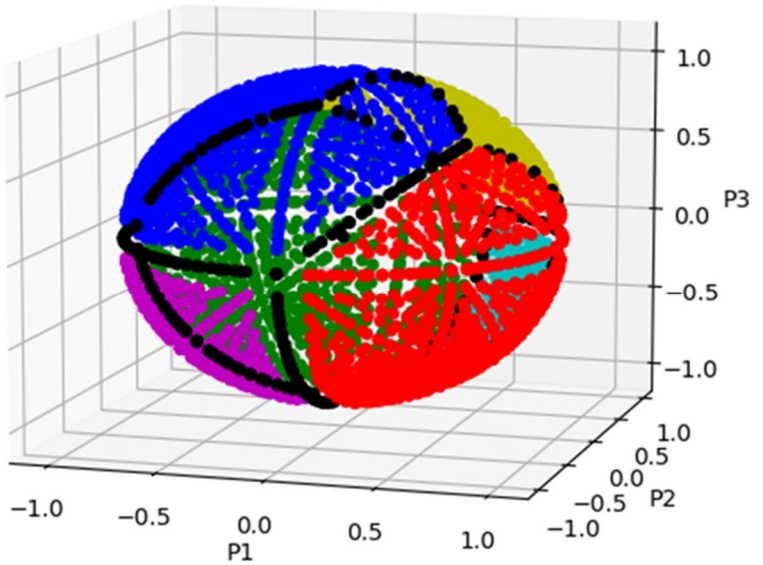

図 4　重みベクトル空間における最適方策分布の可視化

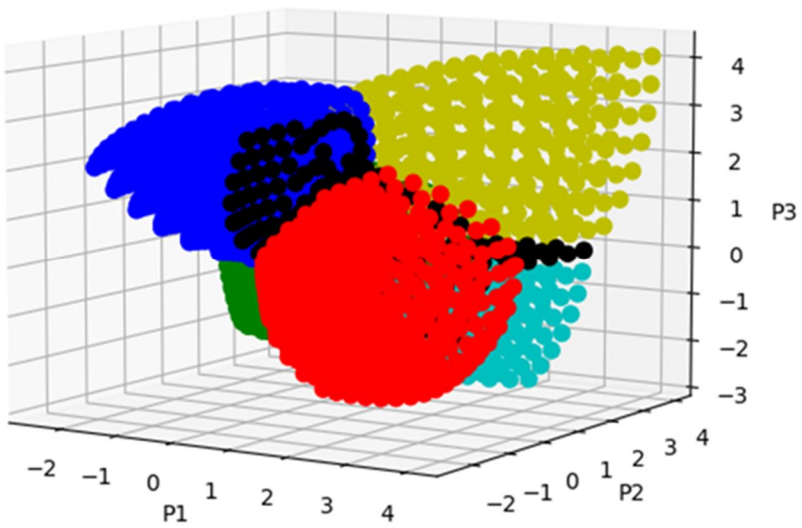

図 5　重みベクトルに対する最適方策の平均報酬$\rho_\pi(\vec{w})$ の可視化

最終年度は，前年度の成果に基づいて，(項目 5)「目的間の任意の重みに対応した多目的最適方策を自動選択する機構の評価」を行った．

まず，(項目 5)のボトルネックについて検討した．ボトルネックは，(項目 4) 多目的最適方策を最適化するための，目的間の重みベクトルの範囲の決定である．メッシュ法での重みベクトル空間の最適範囲を決定する場合，平均報酬算出の計算量は，多目的最適方策数を n，目的あたりのメッシュ点数を m，目的数を k としたときに，o(n x m x k) となる．ここで，m x k は，メッシュの点数(細かさ)を表し，平均報酬の計算に用いる重みベクトル数を表す．つまり，提案手法では，メッシュの各点となる重みベクトルそれぞれについて(項目 1)で収集した多目的最適方策数 n だけ，式(1)を用いて，各方策の平均報酬値を算出し，最大となる方策を決定している．

次に，本手法の優位な点について述べる．従来手法では，平均報酬最大となる多目的最適方策の境界を解析的に解くのが，目的数 3 以上の場合に困難であったのに対し，本手法では，目的数 3 以上の場合について，各重みベクトルに対して，式(1)を用いて各方策の平均報酬値を算出し，最大となる方策を決定するため，計算コストの許す限り，近似的な算出が可能である．しかも，多目的最適方策の決定過程において，多目的間の重要度を表す重みベクトルとは独立な，報酬生起確率ベクトルを(項目 1)でまず算出し，次にそれを用いて(項目 4)で多目的最適方策を最適化するための，目的間の重みベクトルの範囲の決定を，メッシュ法を用いて近似的に行うことで，目的数 3 以上の場合の算出を実現している．

今後の課題は，提案手法の計算コストの削減および状況の変化に応じて多目的方策を動的に制御するための仕組み[Yamaguchi 2023]である．これらを実現するアイデアとして，多目的方策を目的ごとの部分方策(option)に分解し，実行時に状況に応じて各部分方策を動的に組み合わせることによって，多目的最適方策全てを求める計算コストを削減する予定である．具体的には，各目的を学習におけるサブゴールとみなし，各サブゴールを達成する部分方策(option)を本手法を用いて収集，算出することで計算コストを削減する．次に状況の変化を表す文脈をサブゴール系列として定式化し，与えられたサブゴール系列に従って，多目的最適方策を実行時に組合せる手法の実現を目指す．

主な研究成果

[Yamaguchi 2020] Yamaguchi, T., Nagahama, S., Ichikawa, Y., Honma, Y. and Takadama, K.:“Model-Based Multi-Objective Reinforcement Learning by a Reward Occurrence Probability Vector”, in Habib, M.K. (ed.), “Advanced Robotics and Intelligent Automation in Manufacturing”, Chapter 10, pp.269-295, IGI Global, 2020
DOI: 10.4018/978-1-7998-1382-8.ch010
[Yamaguchi 2022] Yamaguchi, T., Kawabuchi, Y., Takahashi, S., Ichikawa, Y. and Takadama, K.:“Formalizing Model-Based Multi-Objective Reinforcement Learning With a Reward Occurrence Probability Vector”, in Habib, M.K.(ed.), “Handbook of Research on New Investigations in Artificial Life, AI, and Machine Learning”, Chapter 12, pp.299-330, IGI Global, 2022.
[Yamaguchi 2023] Yamaguchi, T., Kawabuchi, Y., Ichikawa, Y. and Takadama, K.:“The explainable model to multi-objective reinforcement learning toward an autonomous smart system”, (submitted)

５．主な発表論文等

〔雑誌論文〕　計6件（うち査読付論文　6件／うち国際共著　0件／うちオープンアクセス　1件）

１．著者名 Yamaguchi Tomohiro、Kawabuchi Yuto、Takahashi Shota、Ichikawa Yoshihiro、Takadama Keiki	４．巻 1
２．論文標題 Formalizing Model-Based Multi-Objective Reinforcement Learning With a Reward Occurrence Probability Vector	５．発行年 2022年
３．雑誌名 Handbook of Research on New Investigations in Artificial Life, AI, and Machine Learning, Chapter 12	６．最初と最後の頁 299～330
掲載論文のDOI（デジタルオブジェクト識別子） 10.4018/978-1-7998-8686-0.ch012	査読の有無 有
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 －

１．著者名 Maekawa Yoshimiki、Yamaguchi Tomohiro、Takadama Keiki	４．巻 Vol. 12766, Part II
２．論文標題 Analyzing Early Stage of Forming a Consensus from Viewpoint of Majority/Minority Decision in Online-Barnga	５．発行年 2021年
３．雑誌名 Human Interface and the Management of Information, Lecture Notes in Computer Science	６．最初と最後の頁 269～285
掲載論文のDOI（デジタルオブジェクト識別子） 10.1007/978-3-030-78361-7_20	査読の有無 有
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 －

１．著者名 上野 史, 北島 瑛貴, 高玉 圭樹	４．巻 Vol. 36, No. 6
２．論文標題 複雑ネットワークに基づく多次元意見共有モデル上の誤報伝搬防止	５．発行年 2021年
３．雑誌名 人工知能学会論文誌	６．最初と最後の頁 1～12
掲載論文のDOI（デジタルオブジェクト識別子） なし	査読の有無 有
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 －

１．著者名 Kitajima, E., Murata, A., and Takadama, K.	４．巻 1564
２．論文標題 Multi-value opinion sharing based on information source influence in agent-based network	５．発行年 2020年
３．雑誌名 Journal of Physics: Conference Series	６．最初と最後の頁 1--11
掲載論文のDOI（デジタルオブジェクト識別子） 10.1088/1742-6596/1564/1/012034	査読の有無 有
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 －

１．著者名 Maekawa, Y., Yamaguchi, T., and Takadama, K.	４．巻 1322
２．論文標題 Towards Agent Design for Forming a Consensus Remotely Through an Analysis of Declaration of Intent in Barnga Game	５．発行年 2021年
３．雑誌名 Advances in Intelligent Systems and Computing (AISC)	６．最初と最後の頁 540--546
掲載論文のDOI（デジタルオブジェクト識別子） 10.1007/978-3-030-68017-6_80	査読の有無 有
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 －

１．著者名 Maekawa, Y., Uwano, F., Kitajima, E., and Takadama, K.	４．巻 12185
２．論文標題 How to Emote for Consensus Building in Virtual Communication	５．発行年 2020年
３．雑誌名 Lecture Notes in Computer Science	６．最初と最後の頁 194--205
掲載論文のDOI（デジタルオブジェクト識別子） 10.1007/978-3-030-50017-7_13	査読の有無 有
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 －

〔学会発表〕　計11件（うち招待講演　0件／うち国際学会　2件）

１．発表者名 加藤 駿，速水 陽平，中理 怡恒，高玉 圭樹
２．発表標題 適応範囲の拡大に向けたMAMLとMLSHの組み合わせによるメタ強化学習
３．学会等名 計測自動制御学会，第49回知能システムシンポジウム，2022/3/14
４．発表年 2022年

１．発表者名 戸板 佳祐，前川 裕介，加藤 駿，福本 有季子，中理 怡恒，高玉 圭樹
２．発表標題 他船のモデル化を通した目的地と衝突回避方針の同時推定に基づくマルチエージェント強化学習
３．学会等名 計測自動制御学会，第49回知能システムシンポジウム，2022/3/15
４．発表年 2022年

１．発表者名
福本 有季子，速水 陽平，中理 怡恒，高玉 圭樹

２．発表標題
他エージェントの不確実性にロバストな経路獲得に向けたマルチエージェント逆強化学習

３．学会等名
計測自動制御学会，システム・情報部門 学術講演会 2021 (SSI2021)

４．発表年
2021年

１．発表者名
川端祐也，市川嘉裕，山口智浩

２．発表標題
XAIを用いたノイズに頑健なモデル構築手法の提案

３．学会等名
情報処理学会第84回全国大会，6T-03，2022年3月5日

４．発表年
2022年

１．発表者名
Hayamizu, Y., Amiri, S., Chandan, K., Takadama, K., and Zhang, S.

２．発表標題
Efficient Exploration in Reinforcement Learning Leveraging Automated Planning

３．学会等名
The 3rd Robot Learning Workshop: Grounding Machine Learning Development in the Real World（国際学会）

４．発表年
2020年

１．発表者名
Hayamizu, Y., Amiri, S., Chandan, K., Takadama, K., and Zhang, S.

２．発表標題
Guiding Robot Exploration in Reinforcement Learning via Automated Planning

３．学会等名
The 31st International Conference on Automated Planning and Scheduling (ICAPS 2021)（国際学会）

４．発表年
2021年

１．発表者名
藤本祥，市川嘉裕，山口智浩

２．発表標題
Webページの配色のためのインタラクティブな推薦システムの試作

３．学会等名
情報処理学会第83回全国大会

４．発表年
2021年

１．発表者名
福本大介，市川嘉裕，山口智浩

２．発表標題
テストケース生成補助に基づくプログラミング学習支援

３．学会等名
情報処理学会第83回全国大会

４．発表年
2021年

１．発表者名
山根 大輝，前川 佳幹，荒井 亮太郎，福本 有季子，佐藤 寛之，高玉 圭樹

２．発表標題
正しい意見共有に向けたユーザの投稿頻度を考慮したエージェントネット　ワークシステム：人とエージェントの関係から人とエージェント集団の関係　への展開

３．学会等名
人工知能学会，HAIシンポジウム2021

４．発表年
2021年

１．発表者名
速水 陽平，Zhang Shiqi，高玉 圭樹

２．発表標題
知識の誤りに対する自動計画を利用したモデルベース強化学習のロバスト性

３．学会等名
計測自動制御学会，システム・情報部門 学術講演会 2020（SSI2020）

４．発表年
2020年

１．発表者名 速水 陽平, Amiri Saeid, Chandan Kishan, Zhang Shiqi, 高玉 圭樹
２．発表標題 モデルベース強化学習における自動計画を用いた探索戦略
３．学会等名 情報処理学会, 第19回情報科学技術フォーラム (Forum on Information　Technology: FIT2020)
４．発表年 2020年

〔図書〕　計0件

〔産業財産権〕

〔その他〕

―

６．研究組織

	氏名 （ローマ字氏名） （研究者番号）	所属研究機関・部局・職 （機関番号）	備考
研究分担者	高玉　圭樹 (Takadama Keiki) (20345367)	電気通信大学・大学院情報理工学研究科・教授 (12612)	
研究分担者	市川　嘉裕 (Ichikawa Yoshihiro) (60805159)	奈良工業高等専門学校・情報工学科・助教 (54601)	

７．科研費を使用して開催した国際研究集会

〔国際研究集会〕　計0件

８．本研究に関連して実施した国際共同研究の実施状況

共同研究相手国	相手方研究機関