1F1-02

強化学習とBDIの統合について — カヌー・レーシングを例題とした統合手法の考察 —

高田 司郎[*1] Shiro Takata　山川 宏[*2] Hiroshi Yamakawa　宮崎 和光[*3] Kazuteru Miyazaki　新出 尚之[*4] Naoyuki Nide　長行 康男[*5] Yasuo Nagayuki　酒井 隆道[*6] Takamichi Sakai

[*1]近畿大学理工学部情報学科 Informatics, School of Science and Engineering, Kinki University

[*2](株) 富士通研究所 FUJITSU LABORATORIES LTD.

[*3]大学評価・学位授与機構 National Institution for Academic Degrees and University Evaluation

[*4]奈良女子大学理学部 Faculty of Science, Nara Women's University

[*5]奈良先端科学技術大学院大学 NARA Institute of Science and Technology

[*6]NTT コミュニケーション科学基礎研究所 NTT Communication Science Laboratories

In this paper, we discuss the integration of reinforcement learning and the BDI framework. In general, the former cannot deal with changes of environment, while the latter lacks a learning system but can flexibly responds to changes of environment. By integrating them, we aim at realizing a mechanism that learns nested skills automatically, and intentionally selects a nested reinforcement learning system flexibly from ones it learnt. In particular, using an example of canoe racing, we propose an integrated system of reinforcement learning and BDI for that issue, and study comparisons of it and other learning systems in existence.

1. はじめに

急流をカヌーで流れ下る競技者が滝に出くわした時，その手前で何を考えるだろう．経験から得たスキルやコーチの意見を考慮しつつ，その時のコースコンディションに適合したシミュレーション (熟考) を行い，例えば,「可能な限り左側のほうを行こう，そして二つの大きな岩の間を抜けよう，それから次の岩石群のあたりを後ろ向きに右に行こう.」というように実現できそうなプランを意図して構成する．そしてその後，ついに「えぃ！」とスタートを切るであろう．しかし，実際に滝に飛び出して，カヌーを操る段にいたると，その意図は行為の象徴となり，状況に応じた身体化された反射的なスキル (無意識な行動) の活用が前面に現れることになる．この場面では，スタート前に立てたプランは捨てさられ，物質的・社会的な周辺環境に依存する状況的行為 [4] が，主役に躍り出る．

このように，実世界で課題を遂行するエージェントは，熟考的な能力と，反射的な能力の両方を必要とする．そのため，二つの能力を結合しようとする認知アーキテクチャ研究はこれまでにもあるが，我々は，二つの能力の結合を，意図という一種の心的状態を介して実現しようと試みる．意図は，信念や目標の概念には還元されない，ある程度の時間に渡り維持される未来指向的な心的状態である [1].

意図が，目標指向の行動決定において重要な役割を果たすとする合理的エージェントのフレームワークに，Rao, Singh らによる BDI logic[2] とそれに基づく BDI アーキテクチャ[3] がある (ここでは総称して BDI と呼ぶ)．BDI アーキテクチャは「人は，目標を達成するために，熟考した大まかなプランを意図として形成し，その意図を入力として，動的な周辺環境に依存した最適なプランを，適宜，実践的推論で選択することで，一貫した行為を行っている」という考えに基づく．そのため，ゴールや環境の変化に柔軟に対応できる．反面，設計者がエージェントの行動をプランとして与える必要がある．よって，無意識の反応などはプランとして設計しにくい．

反射的行動の学習に関する研究は，主に強化学習分野でなされている．強化学習エージェントは，問題空間中の報酬関数さえ得られれば，設計者による詳細なプランの設計をしなくても，予測報酬を最大化する行動の自動獲得が可能で，しかも実行速度が速い点で優れている．しかし，環境 (特に，状態に対する報酬関数) が一定であることを前提とするので，後述するカヌー・レーシング問題等のように報酬を割り当てられるゴール (状態) が比較的短時間の間に変化する学習には向かない．

結局，従来の強化学習では，環境が一定であることを前提としており，カヌー・レーシング問題のような目標地点を含め状態が変化する学習には不向きである．一方，BDI アーキテクチャは，プランは与えられていることを前提にしており，この例題のように，周辺環境に応じて無意識に反応するようなプランを記述することは難しい．

そこで，本稿では，強化学習と BDI アーキテクチャを意図を介して統合することで，環境変化に対応できない強化学習の弱点と，学習機能を持たない BDI の弱点を相互に補うことが可能なエージェントの検討を行う．

以下，2 章では，カヌー・レーシング問題を例として，統合に向けた検討課題として，(ボトムアップ課題) 階層的なスキルを自動的に学習する方式と，(トップダウン課題) それら学習した柔軟な階層的強化学習システムを意図的に選択する機構の実現方式，の二つを課題として取り挙げる．3 章では，BDI の立場からの解決方式を，4 章では，強化学習の立場からの解決方式を提案し，5 章では，それら方式を踏まえた強化学習と BDI を統合したエージェントを提案する．6 章では，従来の学習方式や選択方式と比較検討を行い，7 章でまとめる．

2. 統合に向けた検討課題

例題として，カヌー・レーシング問題を説明した後，統合に関わる二つの検討課題を述べる．

連絡先: 高田 司郎，近畿大学理工学部情報学科，〒 577-8502 東大阪市小若江 3 丁目 4 番 1 号，TEL:06-6730-5880，FAX:06-6730-1320，*shiro@info.kindai.ac.jp*

2.1 カヌー・レーシング問題

カヌー・レーシング競技会に素人が臨むことを考える．素人が競技会でよい成績を修めるためには，まず，カヌーの基本的スキルの練習を行う必要がある．本稿では，そのための学習手法として，一種の教師 (コーチ) 付き強化学習手法を用いる．ここで，コーチは，目標地点という形で練習課題を与える．学習者は，その与えられた目標地点を含めた形の政策を学習する．これにより，与えられた任意の目標地点に到達できる能力の獲得を目指す．

基本的スキルが身についた後，カヌーの競技会に臨む．そこでは，ゴールまでの大まかな意図が与えられるという設定の下で，先に学習した目標地点 (への到達方法) と意図との関係を，適宜，調整することでゴールを目指す．これにより，その日，そのときの川の状況に応じた (先に学習した)「基本的スキル」の活用を期待する．

2.2 二つの課題

意図を介して，強化学習と BDI アーキテクチャの統合を行うには，強化学習による意図の利用に関わるボトムアップ側の課題と，BDI 側が意図を与える場合の柔軟性に関わるトップダウン側の課題が現れる．

(ボトムアップ課題) 階層性を持つスキル獲得:

カヌー・レーシング問題に関するスキルは学習の中で階層性が現れてくるものと仮定する．例えば，まず「直進，右に曲がる，左に曲がる」などの基本行為を学習する．次に「ゆっくり直進する，早く直進する，…」などの基本行為の繰返しによる速度に関する学習をする．さらに「目標地点を目指す，障害物を回避する，流れに任す」など目標地点を設定した本稿における基本的スキルを学習する．その後，最終ゴールに到達するために設定されたランドマーク（サブゴール）に到達するための実践的スキルを学習する．そして，これらの階層的なスキルを学習した強化学習システムを選択するスキームが必要になる．このような選択機能が装備されたとして，一旦選択された学習システムは，選択されたゴールを目指して実行される．さて，このような階層的なスキルを自動的に学習するにはどうしたらよいだろうか．

(トップダウン課題) 柔軟な強化学習システムの選択:

カヌー・レーシング問題においては，当初，A という岩を目指してカヌーを漕いでいたが，途中の流れが速く，このままでは岩に衝突してしまいそうなので，当初の目標を捨て，岩の手前 5 メートル地点に目標を動的に変える必要がある．また，局所的には，ほぼ同じパフォーマンスを持っている選択肢がある場合，例えば，距離が遠いが速い流れを利用した経路と，単純に最短距離を進む経路があるような場合に，意図を明確にもたないと中途半端にその間の経路を通ってしまうようなことが起こる．例えば，強化学習において，単純に同じ環境で行うとどちらかの経路に特化してしまうので，レースによって，両方の経路が使えたり，片方しか使えなかったりするため，意図をコントロールできないエージェントは混乱するかもしれない．明確な意図をもって強化学習システムが選択されれば，このような状況は避けられる．このように，従来の階層的強化学習とは異なる柔軟な強化学習システムの選択を意図的に行う機能を課題とする．また，選択機能は，一旦選択した強化学習システムの行為を観察でき，ある許容範囲を持って，成功裏に実行されているか不成功に終わっているかを判断できなければならない．つまり，柔軟に選択した強化学習システムを捨て去る機能も必要である．

3. BDI 側からの統合アプローチ

本章では，BDI のフレームワークについて述べた後，BDI 側からの統合についての検討と提案を行う．

3.1 BDI アーキテクチャ

BDI アーキテクチャ[3] は，動的に変化する環境を知覚し，合理的に問題解決を行うためにプランを選択しながら動作する，熟考型エージェントの内部アーキテクチャである．エージェントの心的状態 (信念・目標・意図) と，プランライブラリ，イベントキュー，インタプリタなどから構成される (図 2)．

心的状態やその時間変化は，それらを陽に表現できる様相論理体系である，BDI logic[2] によって表現される．また，Bratman の意図の理論 [1] で分析されている意図と信念の整合性 (人は，達成の可能性があると信じないことを意図することはない) や，[3] で分析されているコミットメント戦略 (意図が，その達成を信じるまで持続する 'Blind commitment' や，目標が失われてしまうと意図を放棄する 'open-minded commitnent' など) といった，意図の形成に関わる諸条件も，BDI logic で表現される．BDI アーキテクチャは，それらを満たすように，環境の知覚 (信念) と自らの目標からプランを用いて意図を形成し，プリミティブな意図については直接実行，そうでなければサブプランの選択に関する推論を行う．

BDI logic では，未来の複数の可能性を，未来方向に分岐する時間の流れの木で表す．例えば，あるエージェントが，歯を直したいという目標を持つ一方,「歯の治療をすれば歯が痛む」という信念を持つとする．図 1 はこの状況を表したもので (f が「歯を直す」，p が「歯が痛む」を表し，木の根が現在，右方向が未来である)，この場合，Belief world(信念を表す) にない「痛みなく歯を治療する」という未来は Goal world(目標を表す) に現れない．

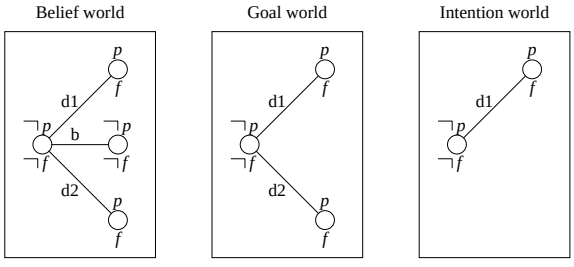

図 1: 歯の治療に関する信念・目標・意図の可能世界

ここでは，意図を表す Intention world は，Goal world の subtree となっている．すなわち，Rao らの分析によれば，エージェントは，望ましい未来のうち (達成の可能性について) 信念と矛盾しないことのみを目標として持ち，さらに，Goal world の時間分岐の中から，自分が達成したい未来を選択して，意図として形成する．そして，Intention world の 1 つ未来の時刻へのどれかの枝に対応する行為を実行する．例えば図 1 の場合，目標が満たされる時間分岐 d1, d2 のうち，d1 を意図として選択しており，この時間分岐に対応する行為が実行される．

3.2 意図の形成と学習 (トップダウン課題への対応)

しかし，プランやルールを用いた推論だけでは，実行する行為を適切に 1 つに絞ることは難しい．現実のエージェントは，限られた時間や計算資源の中で，動的な環境を知覚しつつ意図の形成や行為の選択を行う必要がある．そこで，プランによって形成された意図を達成する複数の行為の候補がありうる場合，その選択を学習によって改善することが考えられる．例え

ば図 1 で，Intention world の時間分岐を d1 だけに絞れない場合，いずれを選択するかを適切な手法によって学習することが考えられる．

2.1 節のカヌー・レーシング問題の例では，川を下った最終目標地点に到達するという目標と，漕ぎ手の知覚，および既に持っているプランのみでは，最善の漕ぎ方の決定はできない．全てを記述しようとするとルールも推論に要する時間も膨大になり，現実的でない．

そこで，当初は中間のいくつかの地点への到達を意図として設定し，流速や障害物に関する知覚情報から，その意図を達成する漕ぎ方に関する，ある程度の制約をルール (プラン) を用いて決め，そこから実際の漕ぎ方の選択を学習で獲得することが考えられる．意図の適切な設定と，プランによる漕ぎ方の制約により，単に全ての漕ぎ方からの選択を学習するより，効率よく，人間の学習課程にも近い学習が行える．

4. 強化学習側からの統合アプローチ

本章では, 強化学習と BDI の統合を検討するために用いるカヌー・レーシング問題の強化学習について述べる．

4.1 基本的スキルの学習 (ボトムアップ課題への対応)

素人がカヌーの基本的スキルを練習することを考える．練習課題はコーチから**目標地点**という形で与えられる．練習者は, その与えられた目標地点を含めた形でスキルの学習を行う．このような学習は, 一般に, 多くの試行錯誤を必要とするため, 強化学習が適している．通常, 強化学習では, センサ入力を状態とし学習を行うが, ここでは, センサ入力に目標地点を含めたものを状態 (**感覚入力**) と呼ぶ．報酬は目標地点に到達した時点でコーチから与えられるが, 目標地点を自ら生成し, 学習することも可能 (自己鍛錬) である．本稿では, カヌー・レーシング問題を中心に議論を進めているので目標地点という用語を用いているが, 一般的には, これは報酬の種類を表すラベル (**報酬ラベル**) と同義である．

目標地点としては, 自分中心な相対座標で与えられる**エゴセントリック**なものと, 世界中心な絶対座標で与えられる**アロセントリック**なものとが考えられる．前者には,「視野内の 3 マス先を目指す」場合などが相当し, 後者には, 海底に固定された (個々に識別可能な) 旗を目指す場合などが相当する．目標地点は,「基本的スキルの学習」においては, トップダウンに与えられるものに過ぎないが, 次に述べる,「実践的スキルの学習」においては, ゴール (最終目標) 達成までの大まかな意図と, 本節で学習した政策とを結びつける重要な要素となる．

いずれにせよ, この時点では, 練習時に知覚した任意のセンサ入力に対し, その知覚を含む感覚入力内に存在するすべての目標地点に適切に到達できることを目指す. 特に, 学習結果が, 目標地点ごとに適切にグループ化されることを期待する．すなわち, 個々のグループには, (そのグループの) 目標地点を達成するために必要なセンサ入力のみが含まれており, その目標を達成する際に無関係なセンサ入力は含まれないことを期待し学習を行う．このような学習は Q-learning や Profit Sharing [7] などの通常の強化学習手法で可能であると考える．

4.2 実践的スキルの学習 (トップダウン課題への対応)

ここでの大目標は試合に勝つことである．この場合, 試合によっては, 様々なコース, すなわち, 様々なゴール (最終目的地) に対応できる必要がある．この場合, まずはじめに, ゴール達成の意図として, 個々の経由地 (**ランドマーク**) を適切に設定する必要がある．試合と同じコース, 同じ状況下での試行錯誤が許されるならば, このランドマークも学習の対象となりえるが, 以下では, ランドマークはコーチから与えられるものとする．なお, ランドマークは, 一般的には, ゴールの達成を手助けするいわゆるサブゴールと同義であり, 4.1 節で述べた報酬ラベルの中のある種の代表であると考えることができる．

個々にランドマークが設定されたとしても, そのランドマークに確実に到達できるかどうかは, どのような練習を行ってきたか, および, その日, そのときの川の状態, 風速の状態等に依存して変化する．この練習とのギャップを, コース前半でキャッチし, それをコース後半で活かすような学習を次に考える．つまり以下では, カヌー・レーシング問題でランドマークが与えられたときに,「そのランドマークに到達するための個々の目標地点の生成」[*1] という意味での意図の具体的な生成方法を述べる．

まず, ここでのエージェントの感覚入力は, 4.1 節での感覚入力から目標地点を除いたもの, すなわち, センサ入力のみとする．エージェントは, はじめに, 現在の目標地点とその目標地点とマッチする過去に練習した (4.1 節で学習した) 学習結果のグループを想起する．そして, そのグループ内の政策のうち, 現在のセンサ入力とマッチする政策に従い行動を出力する[*2]．その結果, 現在の目標地点にグループ化された学習結果にマッチするセンサ入力に遷移できた場合は, その日, そのときの川の状態が, 練習とそんなには違ってない可能性が高いことを意味するので, そのまま感覚-行動サイクルを継続する．一方, 現在の目標地点にグループ化された学習結果にマッチするセンサ入力に遷移できなかった場合は, その日, そのときの川の状態が, 練習とは異なることを意味するので, 出力 (行動) の調整が必要となる．具体的には, 遷移先の状態と本来行きたかった場所との差異をうめるような補正を行う．その補正方法としては, 例えば, 本来 (1,0) を目指していたが,(1,1) に到達してしまった場合には, 目標地点を (1,−1) に補正することなどが考えられる．これにより, その日, そのときの川の状態による補正が, ある程度可能になると期待する．

5. 強化学習と BDI を統合したエージェント

本章では，3, 4 章の検討を踏まえて強化学習と BDI を統合したエージェント (図 2) を提案する．

まず，エージェントは，コーチから目標地点を繰返し与えられて 4.1 節 で述べた基本的スキルを学習する．同時に，その目標地点をゴール，その他のセンサ入力状態を前提条件とするプランを BDI のプランライブラリに何らかの方法で（半）自動登録する．次に，4.2 節 で述べた方式で最終ゴールに到達するために設定された各ランドマーク（サブゴール）に到達するための実践的スキルの学習を行う．同様に，各実践的スキルは 1 つのプランとして，また，コーチが与えたランドマーク列は，最終ゴールを目指すプランとしてプランライブラリに登録する．このような方法で，エージェントは，基本的スキルと実践的スキルを学習しながら関連するプランを登録する．

次に学習後の実践において，エージェントが意図をどのように柔軟に選択し，強化学習システムを選択するかを述べる．まず，エージェントは,「カヌーで優勝すること」という目標を，例えば，コーチから受け取る．そして，option-generator ルーチンにて，この目標を達成するために，上記で登録したランドマーク列のプランを選択して，最終ゴールを目指して優勝するという意図を形成する．次に，deliberate ルーチンにて，今まで形成した意図の中からどの意図を実行するのが合理的かを熟

*1 一般には「サブゴール達成に必要な報酬ラベルの生成」である．
*2 マッチする政策がない場合は, ランダムに行動を選択する．

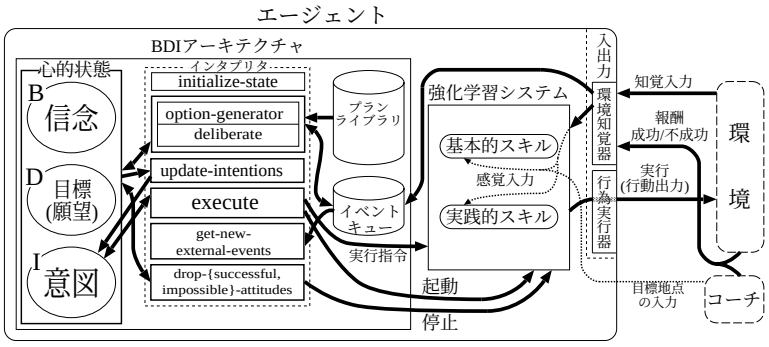

図 2: 強化学習システムと BDI アーキテクチャを統合したエージェント

考する．ここでは，先の意図を選択するとする．そこで, exec ルーチンにて，その意図に書かれている最初のランドマークを意図として選択し，強化学習システムにそのランドマークを目指す実践的スキルを実行するように指令する．

その後，get-new-external-events ルーチンにて，環境情報をイベントキューから取り込み，drop-{successful,impossible}-attitudes ルーチンにて，強化学習システムの行為が成功・不成功かを判断する．不成功，または不成功と予測すると，その起動した強化学習システムを停止する．しかし，そのランドマークを目指すという意図を継続するために，再度，option-generator ルーチンから，その意図を実現する別の実践的スキルを推論し，deliberate ルーチンで意図として選択，exec ルーチンで強化学習システムの起動を繰り返す．強化学習システムは，BDI アーキテクチャから与えられた意図を入力として，学習した基本的スキル，実践的スキルを実行する．このように，我々は，強化学習と BDI の統合システムとして，BDI にて意図の柔軟な選択，強化学習システムにてその意図の実行という意図を介して緩く結合したエージェントを提案する．

6. 考察: 基本的スキル学習の位置付け

今回，BDI アーキテクチャと強化学習の統合を目指して提案した，基本的スキルの学習 (4.1 節参照) の位置づけを行う．そこで，関連技術を含めて，問題空間中の複数の目標地点 (意図に対する行動を生成する能力の比較検討を行う．

強化学習では状態 (行動を含む場合もある) に対する予測報酬により行動を決定する．そこで，従来より，複数の意図に応じた報酬自体を加算することにより複数の目標地点に対応してきた [5]．しかし，この方法では，適宜変化する意図に追随することはできない．既に述べたように，意図には一定時間保持されるものの，適宜変化する性質を持っているためである．これに対して，BDI では，感覚入力 (状態) と区別した意図を明示的に扱い，意図に応じた行動の知識を利用する．

両者の性質を統合する方法として，基本的スキルの学習においては，目標地点 (意図) に応じて行動決定を行うために，強化学習を拡張し，意図を感覚入力の一部として扱うことにした．つまり，“意図” を “報酬を割り当てた状態” とみなすことにより，BDI と強化学習の統合を実現したと言える．

提案手法も強化学習同様，経験から状態に対する予測報酬を蓄積するのに対し，認知距離学習器は問題空間中のある感覚入力 (状態) から任意の感覚入力までの距離を蓄積する．そして，意図までの認知距離が短い行動を選択するように行動を行う [6]．この手法でも意図を明示的に扱うが，同時に意図に複数の状態が存在する場合に適用できない点で異なっている．

7. おわりに

意図を介して強化学習と BDI アーキテクチャを統合するには，強化学習が意図を利用可能とするボトムアップ課題と，BDI が学習により意図を選択するトップダウン課題を解決する必要がある．カヌー・レーシング問題を例として検討を行った結果，目標地点を入力状態に含めた強化学習 (基本的スキル学習) によりボトムアップ課題を解決し，おおまかなプランを制約として学習範囲を狭めつつ，実行時のエラーを修正する局所的な学習を行うことにより，トップダウン課題を解決するという提案を行った．また，基本的スキル学習については，従来の学習方式や選択方式と比較検討を行った．

今後は，カヌー・レーシング問題を実装することで，強化学習システムと BDI との統合に関する検証を行う予定である．

参考文献

[1] Michael E. Bratman. *Intention, Plans, and Practical Reason*. Harvard University Press, 1987. (角脇俊介, 高橋久一郎 (訳), 意図と行為 —合理性, 計画, 実践的推論. 産業図書, 1994).

[2] Anand S. Rao and Michael P. Georgeff. Modeling Rational Agents within a BDI-Architecture. In *Proc. of International Conference on Principles of Knowlegde Representation and Reasoning*, pp. 473–484, 1991.

[3] Munindar P. Singh, Anand S. Rao, and Michael P. Georgeff. Formal Methods in DAI: Logic-Based Representation and Reasoning. In *Multiagent Systems*, pp. 331–376. The MIT Press, 1999.

[4] Lucy A. Suchman. *PLANS AND SITUATED ACTIONS*. Cambridge University Press, 1987. (佐伯 胖 他 (訳), プランと状況的行為人間－機械コミュニケーションの可能性. 産業図書, 1999).

[5] 加藤龍憲, 鈴木昭二, 浅田稔. 複数の報酬による強化学習を用いたサッカーロボットのゴール守備行動の獲得. 第 4 回ロボティクスシンポジア予稿集, pp. 289–294, 1999.

[6] 山川宏, 宮本祐司, 馬場孝之, 岡田浩之. 認知距離学習による問題解決器の実行時探索削減の評価と学習プロセスの解析. 人工知能学会誌, Vol. 17, No. 1, 2002.

[7] 宮崎和光, 小林重信. Profit Sharing の不完全知覚環境下への拡張：PS-r*の提案と評価. 人工知能学会論文誌, Vol. 18, No. 5, pp. 286–296, 2003.