学術・技術論文

自己調整学習メカニズム：オープンエンドな環境で発達するエージェントの自律学習行動原理

星　野　由紀子*　河　本　献　太*　野　田　邦　昭*　佐　部　浩太郎*

Self-Regulation Mechanism: A Principle for Continual Autonomous Learning in Open-Ended Environments

Yukiko Hoshino*, Kenta Kawamoto*, Kuniaki Noda* *and* Kohtaro Sabe*

Continual and autonomous learning are key features for a developmental agent in open-ended environments. This paper presents a mechanism of self-regulated learning to realize them. Considering the fact that learning progresses only when the learner is exposed to appropriate level of uncertainty, we propose that an agent's learning process be guided by the following two metacognitive strategies throughout its development: (a) Switch of behavioral strategies to regulate the level of expected uncertainty, and (b) Switch of learning strategies in accordance with the current subjective uncertainty. With this mechanism, we demonstrate efficient and stable online learning of a maze where only local perception is provided: the agent autonomously explores an environment of significant-scale, and self-develops an internal model that properly describes the hidden structure behind its experience.

Key Words: Self-regulation Mechanism, Self-developmental Agent, Autonomous Agent, Open-ended Environment, Continual Learning, Model Acquisition of Environments with Hidden States

1. は　じ　め　に

1.1 研究の背景

近年，ロボティクスの認知行動制御の研究では，自律発達型のアプローチが盛んになっている [1] [2]．この背景の一つには，従来のロボティクスで行われていた「作り込み」による方式への限界感がある．我々もロボット開発の経験を通してこの問題を強く感じ，「インテリジェンス・ダイナミクス」というアプローチを提案し実行している [3]．その工学的な狙いは，オープンエンドな環境の中で自律的行動を通して自己の身体や外界のモデルを自分の内部に学習・構築し，それを利用しながら合理的に振る舞うエージェントの実現である．ロボットにとっての未知環境でロボット自身が多様に振る舞いながら身体にグラウンドした知識を獲得することで，その行動群は自然さや知能感を備えたものとなる．そのためには数多くの認知行動を記憶できる枠組みが必要であると同時に，環境内での継続的な学習，そしてその学習の結果が次の行動や学習に影響を与えるという動的な学習・行動制御課題に取り組まねばならない．

我々は，この課題に対して予測・計画と動機に着目して取り組んでいる．エージェントが，予測・計画により自分を思いどおりに制御できるように内部モデルを学習し，その内部モデルを用いてタスクを遂行する．また，動機は，計画のための目標を与える目的関数を生成することととらえ，この二つによって行動目標を自律的に決めながら自律発達学習が進むメカニズムを考えている．このような予測・計画できるような内部モデルは，抽象的な意味においてある種の「地図」を作ることである [4]．隠れ状態（直接観測できない状態）が多数存在し，動的に広がり変化していく環境の中で，エージェントが予測・行動するための隠れ状態などを解決した「地図」を作り，その「地図」を利用して目的を達成していく．このとき「地図」作成の過程を目にする人にエージェントの知性を感じさせるには，学習および探索行動の効率も大きなポイントになる．

さて，エージェントの自律発達学習に関する従来研究の代表例に，強化学習 [5] がある．強化学習は基本的に単一の報酬関数に対して，その環境内での最適行動列を学習する問題として定義され，効率よく学習する方法論が検討されている．近年では複雑な構造を広範囲にうまく探索学習するために，内部報酬を取り入れた強化学習の研究も行われてきている [6]．この論文では，全体の状態に対する報酬関数の予測値の差分を導入して探索の幅を広げているが，一般的に広く使えるものであっても報酬関数が事前に与えられる点と，隠れ状態のない環境しか扱っていない点が課題である．

これに対し，従来の一般的な強化学習の枠組みとは異なる自律学習行動の原理を提案し，そこにフォーカスを当てたアプロー

原稿受付　2009 年 10 月 4 日
*ソニー株式会社　システム技術研究所
*System Technologies Laboratories, Sony Corporation
■ 本論文は学術性で評価されました．

チも試みられている．例えば，心理学で言われる「好奇心」という内発動機に従って行動しながら構造を獲得する Intelligent Adaptive Curiosity（IAC）[7] の考えがある．ここで提案された行動戦略は，学習を進捗するための重要な駆動原理となり，自己決定した行動によって得られる経験を学習するという，自律発達の視点を強く打ち出している．しかし，エージェントの行動と環境の相互作用において，既知の「初期状態」へ戻るための強いバイアスがかかっている [8]．現実環境では，新奇な状況を探索し続けると既知の環境から遠く離れたまったく未知の状況に陥ってしまうことがあるが，IAC の枠組み内で提案された指針のみでは，このような状況から自律的に回復し，学習を効率よく進捗させる行動原理としては不十分である．

我々も MINDY というモデルを提案しており [9]，学習が進捗することを基本目的とした学習行動を生成することで，内部モデルを獲得していくことが可能である．心理学のフロー理論 [10] に着想を得て，自己が観測できる変数の可制御性に着目し，その能力が高まるように学習する変数を選ぶという目標設定を行って学習する枠組みを提案した．特に，変数の因果関係に合わせて，自己の獲得したモデルと照らし合わせながらやさしい変数から学習をしていき，結果として階層的な制御モデルを学習する点に主眼をおき，振り子の位置制御課題やロボットの小規模な行動獲得問題でその原理確認を行った．しかし，直接観測できる変数の制御のみを対象としており隠れ変数が扱えないこと，各階層での行動目標の選択がランダム原理に基づいており必ずしも効率的でないこと，などが課題として残っていた．

本論文では，直接観測できない変数の取扱いについてはもちろん，効率のよい自律探索に向けてさらに論を進め，オープンエンドな環境に対する自律発達学習のための行動および学習の新しいメカニズムを提案し，実験を通して検証を行った．

1.2 自律発達エージェントにおける自己調整学習メカニズム

学習者が自らの理解状況を意識し，それに応じて適切に学習戦略を調整すると学習効率が大幅に向上するという事実は，教育心理学の分野で広く知られ，また応用されている．メタ認知に基づいて駆動されるこの種の学習プロセスは「自己調整学習」[11] と呼ばれ，自律発達エージェントの学習において我々の提案するアプローチと共通する点が多い．我々のアーキテクチャの特徴は，自律エージェントが絶えず自らの状況理解の程度を判断し，そのメタ認知に基づいて学習を構成する二つの要素，— データの獲得方法（行動戦略）と学習アルゴリズムの適用方法（学習戦略）のそれぞれを自己調整する，という点にある．

データをただ与えられるだけの学習器と異なり，自ら行動を決定する自律発達エージェントでは，過去の行動観測時系列のみならず，現在の認識状況や，学習器内部の獲得構造に関する情報も利用することで，より効率的な学習を実現できる余地がある．例えば，学習状況に応じて，経験の少ないところや未開拓の領域，予測のあいまいなエリアを重点的に探査するというような行動戦略である．

Fig. 1 (a) に概念的に示すように，認識の不確実性が小さく，予測可能性の高い状況では，これ以上学習すべきものがほとんどないため学習効率は低いが，不確実性が高まるにつれて，学習効率は次第に上昇する．これは行動とその結果より得られる期待情報量が増加するためである．一方，不確実性があまりに高い状況では新しい経験を既存の学習構造のどこにも関連付けることができないため，全体の学習効率は逆に大きく低下する．よって，Fig. 1 (a) の山の左側，すなわち不確実性が比較的小さい状況では不確実性を大きくするような探索行動（右上向きの矢印）が有効であり，山の右側，つまり不確実性の大きな状況では不確実性を速やかに低減させる同定行動（左上向きの矢印）が学習効率の向上に寄与する．Fig. 1 (b) は先に述べたフロー理論を簡略化して表した図である．人はスキルとチャレンジがつりあったフロー状態（網掛部）において最高のパフォーマンスを発揮し，またそのような状況を強く追い求めるものとされている．ここで，「自己のスキル」に対する「課題の難易度」の比が我々の指標とする「不確実性」であると見て，Fig. 1 (a) のグラフを重ね合わせてみると，フロー状態というのは学習効率の山の部分に相当し，学習効率を高めるような行動戦略はフロー状態に入るための戦略そのものであると考えられる．

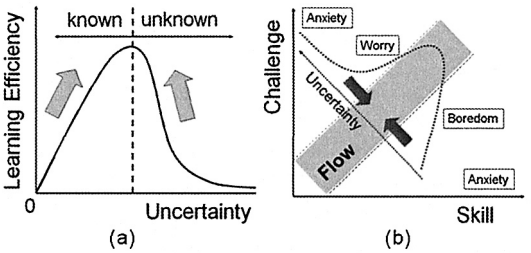


Fig. 1 Rough sketches of uncertainty regulating behaviors (a) Relation between uncertainty and learning efficiency (b) Uncertainty-axis mapped on the skill-challenge plane of the Flow theory

次に，学習戦略について検討する．発達エージェントの学習においてオンラインの追加学習能力は必須である．新しい経験の学習は，その経験を得た状況の理解に基づき統合的に行われるため，学習の基盤となる状況理解の程度を正しく把握すること — メタ認知はここでも極めて重要である．殊に，系の状態が直接観測できない場合は，新しい経験をこれまでに獲得したモデルの中にどう位置づけるべきかが自明ではないため，既存の構造を調整して対応するのか，それとも構造自体を拡張して対応するのかの判断が難しい．この判断を誤るとこれまでに学習してあった既存構造まで崩壊してしまうことがある．

我々は，ここにもメタ認知的制御の考えを応用し，状況認識の不確実性に応じた学習戦略の切り替えを行う．具体的には，まだ十分把握できていない新奇な状況では，局所的な時系列変化を新規モジュールに保持する程度の学習にとどめ，未知の状況から，よく知った既知の状況へと復帰したそのタイミングではじめて，新規モジュールも含めた全構造を改めて全体的に調整しながら追加学習を行うというものである．この学習には，「既知から未知へ，そしてまた既知へ」という一連のシーケンス全体を利用する．この結果，新規モジュールの無駄な生成を抑えつつも既存モジュールの誤学習を防止し，オープンな環境で新規の経験を安定的に獲得することができるようになった．

以上をまとめると，我々の提案する自己調整学習メカニズム：メタ認知的知覚に基づく行動戦略・学習戦略の自律的調整，の概

Table 1 Outline of our self-regulation mechanism. (a), (b), (c), and (d) represents different research areas. Quoted terms are the names of our strategy, described in sections 2. 3 and 2. 4. Meta-strategies for regulating them are not explicitly shown in the table, but explained in the text

	Metacognitive awareness	
	Known	Unknown
Behavioral strategy	(a) 'Exploration'	(b) 'Identification'
Learning strategy	(c) 'Global update'	(d) 'Local update'

要は **Table 1** に示すようなものとなる．(a)「既知」状態における行動戦略，つまり不確実領域への探索行動については，強化学習 [6] や内発動機に基づく自律発達 [8] などの研究として，多くの報告がなされているが，系の状態が直接観測でき，「未知」状態が存在しないような設定であったり，系全体が初期状態への強い復元力を持っていて，「既知」状態での戦略のみが支配的となるような問題であることが多く，「未知」状態での性能も含めて検討されているものは見当たらない．(b)「未知」状態における行動戦略，つまり同定行動に関する研究 [12] もあるが，自律発達学習の文脈で論ずるものはあまり多くない．(c)，(d) に相当する，学習戦略のメタ認知的調節に関する研究報告は我々の知る限りほとんどない．本論文では，(a)，(b)，(c)，(d) すべての領域をカバーするとともに，メタ認知的な観点に基づく，これらの戦略の同期的な切り替えとその効果について論ずる．

以下，2 章では，本論文で提案する自律発達学習メカニズムを，様々なタスク，環境，学習モデルに適用可能なように高いレベルの抽象度で定式化する．3 章では具体的な応用例として，部分観測マルコフ決定過程（POMDP）を取り上げて，そのモデル化に用いる隠れマルコフモデル（HMM）[13] の拡張と，より詳細レベルでの提案手法の定式化を行う．提案するアプローチの有効性を検証する実験とその結果は 4 章で報告する．そして，5 章では本実験から得られた知見に対する議論を行い，最後に，6 章で結論と今後の課題に関して述べる．なお，実験時の実装の詳細については付録にまとめてある．

2. 自己調整学習の定式化

これ以降の詳細な議論を進めるにあたり，エージェントおよび環境よりなる系を離散状態マルコフ過程としてモデル化し，次に示す記号を用いて定式化する．

まず，時刻 t におけるエージェントの観測ベクトルとアクションベクトルをそれぞれ $\boldsymbol{o}_t$，$\boldsymbol{u}_t$ と表記し，それらの時系列シーケンスは $\mathcal{O}$，$\mathcal{U}$ と略記する．次に，系の状態をユニークに記述する隠れ状態変数（学習モデルの内部状態変数）を Z とし，時刻 t での具体的な変数値は z_t で表記する．π_t は時刻 t における Z の事前確率分布である．このようなマルコフ過程を記述するモデルはいくつか考えられるが，ここではその詳細を特定せず，モデルパラメータを λ によって抽象化して表現する（3 章では，拡張された HMM に基づくより具体的な実装例を示す）．

現在獲得されているモデル λ と過去の経験 $\mathcal{O}$，$\mathcal{U}$ を基に，よりよいモデル λ' を効率的に獲得するため，エージェントはオープンエンド環境でどう振る舞い，どう学習するべきか，という問いに答えることが自己調整学習メカニズムの目標である．そのために必要となる各ステップについて説明する．

2. 1 系の状態 z_t の推定

時刻 t において，系がどのような隠れ状態にあるのかを推定するためには，事後確率 $P(Z_t|\pi_t, \boldsymbol{o}_t, \lambda)$ を評価すればよい．離散状態マルコフ過程では，モデルパラメータ λ と過去の行動観測シーケンス $\mathcal{U} \equiv \{\ldots, \boldsymbol{u}_{t-1}\}$，$\mathcal{O} \equiv \{\ldots, \boldsymbol{o}_{t-1}\}$ から時刻 t での事前確率 π_t，そして事後確率 $P(Z_t)$ を推定することができるが，ここで問題となるのは，どれだけ過去まで遡ったシーケンスデータを利用するか，である．我々は 付録 A 章に示すとおり，認識ウィンドウ長 w を状況に応じて動的に変化させることでこの問題に対処する．この結果，時刻 $t-w$ から t までの各時刻における内部状態 Z の確率分布時系列 $\{P(Z_{t-w}), \ldots, P(Z_t)\}$ および最尤状態時系列 $\{\hat{z}_{t-w}, \ldots, \hat{z}_t\}$ が計算できる．

2. 2 「既知」および「未知」のメタ認知的知覚

現在の状況が学習されたモデルでどの程度うまく説明されるのか，つまり自らの状況理解の程度が十分なのかそうでないのかを判断するメタ認知的知覚の過程は自己調整学習の第一歩である．これを実現するための基本的な考え方は次のようになる．(1) 前節で述べた可変ウィンドウ長認識において，ウィンドウ長の変動に対して安定的な認識結果が得られ，(2) さらに状況が一意に特定できる，つまり，エントロピー $H(P(Z))$ が十分小さいこと．(3) その上，得られた最尤推定時系列 $\{\hat{z}_t\}$ が経験に照らして突飛なものではないならば，現状は十分よく理解された「既知」状態であると考える．逆にいずれかの条件が満たされないならば，既存のモデルではうまく説明できない「未知」状態である可能性が高い（より具体的には 付録 B 章を参照）．

このように，自らの状況理解の程度を評価することで，単に最尤状態系列を信じて行動するのではなく，現状を「未知」として保留したまま，そのメタ認知に応じた別の行動・学習戦略をとることが可能となる．このとき，採用される行動・学習戦略は状態推定 Z_t の一意性・単峰性を仮定せず，幅広い可能性を許容するものでなくてはならない．

2. 3 行動戦略のメタ認知的調節：探索戦略と同定戦略

学習を効率よく行うためには，状況の不確実性をある適切なレベルに保つことが重要である．不確実性なくすべてが予見できるような状況ではこれ以上学習すべきものは何もないが，その一方で，あまりに不確実性の高い状況では有効な学習がほとんど不可能となってしまう [7]．この節では，不確実性を増加させる「探索」戦略と減少させる「同定」戦略をうまく切り替えて，高い学習効率を維持するための方法を説明する．

2. 3. 1 探索：不確実性を増加させるための戦略

これは，現在の事前確率 π_t および観測 $\boldsymbol{o}_t$ のもとで，これ以降のある時刻におけるエントロピー $H(P(Z))$ の期待値を最大化するようなアクションシーケンス $\mathcal{U} \equiv \{\boldsymbol{u}_t, \ldots\}$ を生成するものである．将来起こりうるすべての観測時系列の集合を $\mathcal{O}$ として，この戦略は $\underset{\mathcal{U}}{\operatorname{argmax}} \sum_{\mathcal{O}} P(\mathcal{O}|\pi_t, \boldsymbol{o}_t, \mathcal{U}, \lambda) H(P(Z|\pi_t, \boldsymbol{o}_t, \mathcal{O}, \mathcal{U}, \lambda))$ と記述できる（詳細は 付録 C 章を参照）．

ここで $\mathcal{U}$ に関して最大化されるのは，将来起こりうるすべて

の時系列 $\mathcal{U}$，$\mathcal{O}$ のもとでの状態事後確率のエントロピーの期待値である．したがって，この戦略は未経験の遷移を探索するようなアクション $\mathcal{U}$ を生み出すことになる．

2.3.2 同定：不確実性を減少させるための戦略

この戦略は探索戦略のちょうど反対となるように定式化され，$\mathrm{argmin}_{\mathcal{U}} \sum_{\mathcal{O}} P(\mathcal{O}|\pi_t, \boldsymbol{o}_t, \mathcal{U}, \lambda) H(P(Z|\pi_t, \boldsymbol{o}_t, \mathcal{O}, \mathcal{U}, \lambda))$ と表せる（詳細は 付録 C 章を参照）．

これは内部状態 Z の不確実性をできるだけ減少させる行動を導くものであるが，新たな環境変化が起きたり未探索領域に踏み込んでいるときなど，環境の不確実性が非常に大きい場合，π_t の推定が困難になるという問題を抱えている．これに対処するため，我々は 2.1 節で述べた可変長認識ウィンドウに基づく内部状態推定を応用する．ただし，付録 A 章 ステップ（3）でのエントロピー系列の収束判定に代えて，付録 B 章 ステップ（3）で説明した行動状態系列の生起確率判定を用いる．この結果，π_t はモデルに整合する範囲でできるだけ過去まで遡ったデータシーケンスを用いて推定されることになるので，過去に経験して獲得された記憶構造のうち，直近の経験に適合する可能性のある部分すべてを考慮した行動計画が可能になる．

2.3.3 探索と同定を切り替えるためのメタレベル戦略

基本方針は極めてシンプルであり，「既知」状態では「探索戦略」を，「未知」状態では「同定戦略」を採る．この結果，学習が十分進み状況の不確実性が少なくなったときには不確実性の多い状況へ移動するような行動が現れ，学習がまったくなされていない未知領域では逆に，現状をできるだけ早く同定して不確実性の少ない状況へ戻ろうとするような行動が現れるので，不確実性のレベルが適切に制御され，エージェントの学習効率も高く保たれる．このように未知領域と既知領域を交互に行ったり来たりするような行動は次に述べる学習戦略の観点からも非常に重要な意味を持つ．

2.4 学習戦略のメタ認知的調節：全体学習と局所学習

エージェントがオープンエンドな環境で自律的に行動し，新たな経験を獲得して学習するためには追加学習の機能が必須である．では，エージェントはいつ追加学習を行うべきであろうか．これまであまり指摘されていないことだが，仮に学習データが同じであっても追加学習のタイミングは学習結果に大きな影響を及ぼす．具体的には 4.3 節で示すが，我々の経験によれば，機械的に追加学習プロセスを走らせるだけでは不十分であるどころか，むしろ学習結果を破壊することすらある．エージェントは自らのメタ認知に基づき，学習戦略を適応的に変化させなければならないのである．

2.4.1 全体学習戦略

これは通常の追加学習プロセスそのものである．最近経験した行動観測シーケンスに基づき，モデルのパラメータ λ が（それ以前の経験も加味しつつ）適切に更新される．多くの場合，行動観測シーケンスの尤度または事後確率を最大化するための最適化計算が行われることになる．学習対象となる行動観測シーケンス $\mathcal{U}$，$\mathcal{O}$ が，現在のモデルである程度正しく理解され記述できるのであれば，この追加学習プロセスは有用であり，モデルを洗練するのに役立つ．

2.4.2 局所学習戦略

ここでの学習は局所的，場当たり的なものである．これまでのモデルでは記述しきれないまったく新しい状況での新しい経験に対応するため，モデルが局所的に拡張され，即時学習的に最近の経験が格納される．全体とは切り離された局所領域のみが最近の経験をうまく説明するように追加・更新され，「既知」状態からの遷移によってゆるやかに全体との関連性を持つので，モデルパラメータのほとんどは変更されないまま保たれる．「未知」状態を記述するための新しい内部状態 z_t が次々と追加される一方で，それら同士の関係は時間的に連続するものが局所的に定義されるのみであるため，これらの暫定的な状態遷移構造は本質的に疎である．これを既存のモデルに接続するのは，「既知」状態から「未知」状態，また「未知」状態から「既知」状態へと遷移する瞬間の行動観測シーケンスとなる．

2.4.3 全体学習と局所学習を切り替えるためのメタレベル戦略

「既知」状態においては，追加学習が絶えず行われ（全体学習戦略），モデルはどんどん洗練されていく．一方，既存モデルではうまく説明できない「未知」状態になると，全体学習，つまり通常の追加学習プロセスは直ちに停止され，代わって局所学習が行われるようになる．同時に，「未知」状態におけるすべての行動観測シーケンスが記録されはじめる．同定戦略に基づき環境が調査され，最終的にすべての状況がうまく説明できるようになると，メタ認知的知覚も「未知」状態から「既知」状態へ変化するが，その際に，この間に記録されたすべてのデータを用いた全体学習が行われ，局所学習時に追加された内部表現もモデル全体に融合される．

3. HMM を用いた POMDP 環境のモデリングとそれに基づく自己調整学習の実装

これまでの議論は，特定のタスク，環境，学習モデルに依存しない形式で行ってきたが，それは，この自己調整学習メカニズムが自律発達エージェント全体に対して適用できる一般性の高いフレームワークであることを示すためだった．これ以降，このフレームワークをより具体的なタスク，環境，そして学習モデルに対して適用していく．なお，全体のシステム構成やエージェントの動作フローについては 3.6 節で述べる．

自律発達エージェントの研究において避けて通ることのできない隠れ状態を取り扱うため POMDP による定式化を考えるが，本論文では環境からの外部報酬を陽に取り扱わない．環境からの報酬がなくとも，自分だけで内部的に目標を作り出して探索を駆動する自律発達エージェントが我々の目標だからである．同時に，報酬の存在によって環境中の重要な構造に対するヒントを与えてしまうことを避ける意図もある．

学習モデルとしては，次章で説明するとおり，時系列データの教師なし構造化能力に優れた HMM の拡張形式を用いる．このモデルによって環境の確率構造が獲得されれば，動的計画法を利用した再帰計算により，その環境で許される任意の遷移を効率的にコントロールできるようになる．

本論文では，POMDP のモデルを変数（Z，π，A，B，U，V）によって表現する．ここで，Z は HMM の内部状態，π は

各状態の事前確率，U はアクションの種類である．A はアクション U に依存する状態遷移確率を表し，ある状態 Z_t の下でアクション U を取った場合の，次の時刻での状態 Z_{t+1} の確率分布を与える．B は，状態 Z において観測シンボル V が観測される確率を表す．HMM による定式化では，観測シンボル V のみが直接観測可能で，実際のモデルを構成している隠れ状態 Z は直接観測できない．なお，オープンエンドな環境で自律発達するエージェントを想定しているので，隠れ状態 Z のサイズは実行時に随時増大する．

3.1 アクションに応じた予測モデルとしての HMM の拡張

本研究では， POMDP 環境を予測モデルとして表現するために HMM の拡張を行う．具体的には，エルゴディック（全結合型）HMM [4] に対してアクション表現の拡張を行う．つまり，通常の HMM の持つ，サイズ $Z \times Z$ の二次元状態遷移確率テーブルを，サイズ $Z \times Z \times U$ の三次元状態遷移確率テーブルに拡張して用いる．このアクション拡張型の遷移確率テーブルは，アクション $\boldsymbol{u} \in U$ のそれぞれに対して，その結果引き起こされる状態遷移を表現する 二次元の確率テーブルを個別に持つようなものである．なお，モデルパラメータの推定は，Baum-Welch アルゴリズム [14] に対してアクションの条件付けに関する拡張を行うことにより実現されている．端的には，状態 i から状態 j への遷移確率 a_{ij} の代わりに，その時刻 t におけるアクション $\boldsymbol{u}_t$ のもとでの状態遷移確率 $a_{ij}(\boldsymbol{u}_t)$ を用いる（さらなる詳細については文献 [15] を参照）．

エルゴディック HMM は全状態間で任意の結合が可能であり高い表現能力を持つが，そのパラメータ自由度のゆえに学習が局所解に陥りやすいという難点がある．我々はエルゴディック HMM の学習安定性を高めるために，スプリット・マージ法（付録 D 章）を学習に導入する．これは解に対する構造化制約（一状態・一観測制約）を導入することで，この問題の解決を試みたものである．

3.2 拡張型 HMM におけるメタ認知的知覚の実装

実行ステップごとに，状態推定モジュールは，それまでの観測シンボルとアクションシンボルの系列から，エージェントが環境中の「既知」状態，すなわちすでに経験ずみでモデルとして獲得された状態にいるのか，それともまだ経験したことのない「未知」状態にいるのかの認識を行う．2.1 節で説明した状態認識処理は今回のモデルに対してもそのまま適用できる．内部状態系列の推定には Viterbi アルゴリズムを用いるが，その際には 3.1 節と同様の変更を加える（再掲すると，状態 i から状態 j への遷移確率 a_{ij} の代わりに，その時刻 t におけるアクション $\boldsymbol{u}_t$ のもとでの状態遷移確率 $a_{ij}(\boldsymbol{u}_t)$ を用いる）．付録 B 章ステップ 3 に記述されている確率計算には，観測確率 B および 状態遷移確率 A を用いる．

以上のメカニズムによって，可変長の観測・アクションシンボル系列を入力情報とした，状態推定メカニズムが実現できる．

3.3 探索行動の実装：オープン端探索

エージェントが「既知」状態にある場合，2.3 節で述べたとおり，内部状態の不確実性を増すような行動が行われる．これは概念的には次のように実現される：(1) 比較的探索されていない状態群を見つける，(2) その状態の一つに向かってパスをプランする，(3) そこに行き，経験が少ない行動の一つを試す．

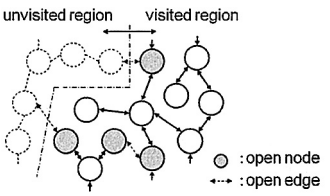


Fig. 2 The definition of open edge and open node

Fig. 2 は未探索領域を表す図である．各ノードは HMM の内部状態を表し，ノード間の矢印は状態間の主要な遷移を表す．点線の円は未発見でまだ構造化されていない状態である．点線の矢印は，「オープンエッジ」と呼ばれ，経験が少ないために未発見の遷移を表す．一つ以上のオープンエッジを持つノードを「オープンノード」と呼び，オープンノードとオープンエッジをまとめて「オープン端」と表現する．

状態 z においてアクション $\boldsymbol{u}$ の遷移先確率が多数の状態に広がった分布をもつとき，(z, $\boldsymbol{u}$) はオープン端と特定できるので，HMM の遷移行列 A を調べれば簡単にオープン端を見つけることができる．オープン端が見つかれば，現在状態からそこに到達するプランは動的計画法により定まるので，オープンエッジを探索するアクション $\boldsymbol{u}$ をその最後に出力すればよい．

この「オープン端探索」の行動戦略により，エージェントは，まだ未知の領域や未経験の遷移を優先して，効率よく環境を動きまわり，学習に有益な情報を集めることができる．

3.4 同定行動の実装

エージェントは「未知」状態にあると判断すると，内部状態 Z の事後確率に関するエントロピーすなわち，不確実性を減少させるための行動戦略に切り替える．この戦略の実現には，状態の事後確率分布に関するエントロピーの期待値を最小化するアクションシーケンスの探索を行う必要がある．アクションシーケンスの探索には，A^* などのグラフ探索アルゴリズムが利用できる．今回は，先読み深さの打ち切りに関する閾値をもつ，幅優先探索を以下の近似とともに用いた．(1) その状態における観測確率が極端に低い観測についてはその可能性を無視する（枝刈り）．(2) エントロピーの最小値を厳密に求めるのではなく，一定の閾値を下回るエントロピーが達成された場合，それ以上の深さへの探索を中止し，同一深さ内でのエントロピー最小状態を目標に定める．(アルゴリズムの詳細は 付録 C 章を参照.)

この結果，エージェントは「未知」環境においても無駄な動きをすることなく，人から見ても自然で知的な振る舞いを見せる．一見すると，「未知」の状況でランダム以上の行動は出せそうもないように感じられるが，実はそうではない．「未知」の状況はいつでも，過去に経験した類似の状況と一致する可能性に開かれており，その可能性を考慮しつつ最適な同定行動を行うことで，一貫性のある効率的な行動が生み出されるからである．それに加え，過去訪れたことがある場所へ（たまたま）戻ってきたときにその可能性を速やかに検出して同定できる効果も大きい．4.2.1 項において，これらの定量的な評価を示す．

3.5 全体学習と局所学習の実装およびその切り換え

エージェントが「既知」状態にある場合，学習モジュールは直近のアクション観測シーケンスを用いて HMM の継続的追加学習を行う．今回用いたアルゴリズムは HMM の集団学習的アプローチを追加学習に応用したもので，文献 [16] において “Ensemble Training for HMM within an Incremental Learning setting” と呼ばれているものに相当する．なお，この学習の際にも付録 D 章のスプリット・マージ法を適用する．

一方，エージェントが「未知」状態にあるとき，学習モジュールは，一つのアクション $\boldsymbol{u}$ を実行して一つの観測シンボル $\boldsymbol{o}$ を得るたびごとに，その観測シンボル $\boldsymbol{o}$ に強く対応付けられた新しい状態をモデル中に作成し，一つ前の状態との間をアクション $\boldsymbol{u}$ に関連付けられた遷移によって接続する．新規状態に関するその他の遷移確率は次のようにして求めた確率で初期化しておく．(1) 1 ステップのアクション観測シーケンス（$\boldsymbol{u}$，$\boldsymbol{o}$）に対する認識を行い，過去の経験の中から類似した状態の候補（認識後の状態確率が高いもの）を選びだす．(2) それらの候補において，自己遷移や一つ前の状態へ戻る遷移など共通の性質を持つ確率構造があれば，その平均確率をあてはめる．特に共通の構造がない場合は一様分布で初期化する．さらに，この期間のアクション観測シーケンスは全ステップを漏らさず記録し続けておく．

このようにして，局所学習を継続しながら 3.4 節の同定行動を続け，メタ認知的知覚が「未知」状態から再び「既知」状態に戻ったなら，「未知」状態で最後に追加された内部状態をその次の「既知」状態へ，対応するアクションに関連付けられた遷移によってこれまでと同様に接続する．その後，記録してあった「未知」状態での全アクション観測シーケンスを用いて HMM を追加学習し，新規の経験を既存のモデルと整合・定着させる．スプリット・マージ法の適用についても同様である．この際，「未知」状態で局所的に追加された内部状態のうち，ループ状の箇所を通って二重に追加された内部状態や，「既知」領域を通っていたにもかかわらず「未知」状態から遷移したことにより「未知」領域として追加された内部状態などが存在するので，内部状態の融合が正しく行われることが重要になる．正しい更新は，既存モデルとの接続の情報が得られて初めて可能になるため，曖昧な箇所は「未知」として状態を追加しておき，あとで全体学習を通して同じ状態を抽出して融合する，という戦略をとる．

3.6 システムの構成と動作

エージェントは，各実行ステップごとに環境に対して一つのアクションシンボルを出力し，状態変化を起こす．状態変化の結果としては，変化後のエージェントおよび環境の状態を反映した観測シンボルが入力される．このように，今回の問題設定では，エージェントと環境の相互作用は，アクションシンボルと観測シンボルの入出力として抽象化されている．**Fig. 3** に，今回の実験系と，エージェントが環境中で行動生成するための全体のシステム構成を示す．

提案手法における実際のエージェントの動きを擬似コードで表すと以下のようになる．なお，擬似コード内のオブジェクト表記（太字）は Fig. 3 内の各オブジェクトに相当し，ドット以下がそのオブジェクトが実行するメソッドを表している．

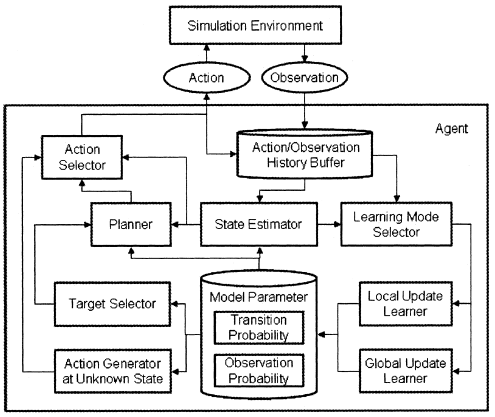


Fig. 3 Block diagram of a self-developmental agent

```
while(1)
  get_current_observation()
  state_estimator.get_current_state()
  if current_state == Known
    if previous_state == Unknown
      local_update_learner.attach_to_global()
    end
    global_update_learner.do_learning()

    if there_is_no_plan
      target_selector.find_open_end()
      planner.make_plan()
    end
    action_selector.pop_from_plan()
  else if current_state == Unknown
    local_update_learner.do_learning()
    action_generator.make_identifying_action()
  end
  action_selector.execute_next_action()
end
```

この擬似コードのように動作することで，「未知」領域と「既知」領域を行き来し，環境と相互作用しながら環境モデルを広げていく．「既知」状態では不確実性の高い方向へ動き，「未知」状態では同定行動を行いつつ学習戦略も切り替える．特に，「未知」状態から「既知」状態に戻ったところで局所的に追加していた領域を既存モデルに接続して，全体学習を行うことで，適切な分離・融合とともに内部モデルが更新される．

4. 自己調整学習メカニズムの有効性検証実験

4.1 実験設定

観測列の背後にある隠れた構造を内部モデルとして獲得する問題を解くために，**Fig. 4** に示すような迷路状の環境とエージェント中心のローカルセンサを用意した．

エージェントは，場所に合わせて Fig. 4 (d) に示すような 16 個の観測シンボルのうちの一つを観測するが，アクションシンボル・観測シンボル相互の関係性について事前知識を持たない．これは，ローカルセンサ情報と自分の出力したアクションだけから，隠れ状態を含む構造を作り出し，さらにそれを次の行動決定に利用できるかどうかを検証するためである．エージェントは Fig. 4 (c) にあるアクションを一つ選んで実行する．エージェントが壁に当たった場合，エージェントの位置は変化しない．2.2 節で述べたメタ認知のためのパラメータは以下のものを使った．$[\theta_H = 1.0,\quad \theta_{P_o} = 0.8,\quad \theta_{P_u} = 0.1,\quad \theta_w = 10]$.

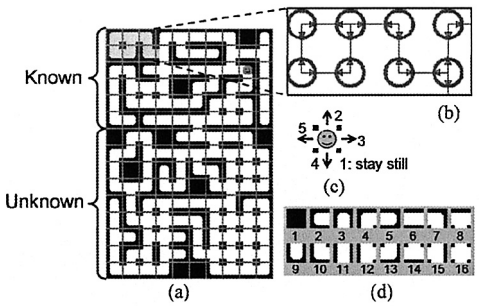


Fig. 4 (a) Maze-like environment (b) Internal states of the learned HMM, arranged using the position information which is hidden to the agent (Circle: internal state. Arrow: transition between internal states) (c) 5 action symbols (d) 16 observation symbols

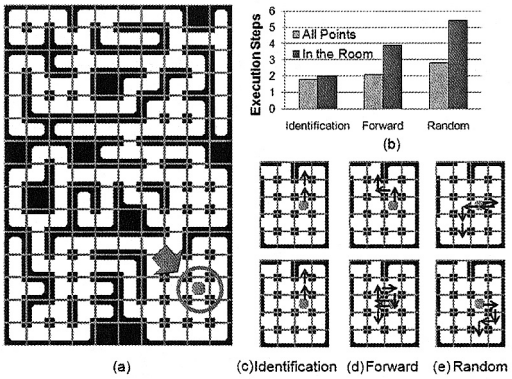


Fig. 5 (a) The world for evaluation (b) Total execution steps needed for identification (c) Action examples in our proposed identification strategy (d) Action examples in 'forward' strategy (e) Action examples in 'random' strategy

各実験では，学習で獲得した内部モデルの再利用性などを検証するため，あらかじめ小さな領域を HMM で初期学習した状態から実験をはじめている．Fig. 4 (a) の中のサイズ 6×10 の 'Known' 領域における経験から 16,000 ステップの観測およびアクション列を作り，8×12 の状態空間をもつエルゴディックHMM でスプリット・マージ法も使いながら学習を行った．この初期学習ずみの HMM を用いて，エージェントは自己調整学習のメカニズムに基づき行動生成を開始する．

4.2 探索行動および同定行動の効果の検証

4.2.1 同定行動戦略の効果

先に説明した探索行動は，オープン端の発見とプランニングの組み合わせなので直感的に動きを想像しやすい．それに対し，未知状態での同定行動がどのような挙動を示すものなのか，式だけからは想像しにくい．そこで，まずはじめに同定行動を用いた場合の実際の動きに関する検証結果を報告する．

Fig. 5 (a) に示す環境で，環境モデルを学習したエージェントを用いて，エージェントを任意の場所におき，そこから自己位置を同定する際の行動出力を比較した．比較対象としては，あと戻りしない行動（「前進行動」と名づける）と移動できるアクションからランダムに選択するもの（「ランダム行動」と名づける）を用いた．「ランダム行動」は実行可能な行動をランダムに選択するのではなく，壁の方向に進むなどの無意味な行動を取り除いた後でのランダムな行動であることに注意して欲しい．本当にランダムな行動では比較にすらならない．

まず，同定にかかるまでのステップ数を 1,000 回試行して平均をとった．結果を Fig. 5 (b) に示す．エージェントをランダムに配置し位置同定を行わせた場合（Fig. 5 (b) All Points），提案手法が 1.8 回，「前進行動」が 2.1 回，「ランダム行動」が 2.8 回となり，提案手法が一番少ないステップ数となっただけでなく，場所によるばらつきも少ないことが分かった．例えば，Fig. 5 (a) の矢印で示す場所の同定を 1,000 回行って平均をとると（Fig. 5 (b) In the Room），提案手法が 2.0 回，「前進行動」が 3.9 回，「ランダム行動」が 5.4 回となり，他の手法との差が大きく広がることが分かる．これは，「前進行動」では T 字路や十字路などでランダム選択が発生するが，提案手法では必ず期待値的に最適な方向へ進むからである．

Fig. 5 (c) (d) (e) には，(a) の位置にエージェントが置かれた場合のそれぞれの行動戦略で実際に出力された行動例を示す．同じ場所から各行動戦略に従って行動し，自己位置を同定するまでの過程を矢印で示し，2 回分を提示している．我々の提案手法 (c) では，どの試行でも安定して短いステップで同定できたが，(d) の「前進行動」では，結果的に遠回りや最適でない方向への行動が出力されたり，(e) に示した「ランダム行動」の行動出力では，往復が起こったりした．たまたま「ランダム行動」や「前進行動」で最短ステップで同定できる行動が選ばれることもあるが，その確率は高くない．

以上により，我々の提案，実装した行動戦略が，効率のよい同定行動として機能していることを示すことができた．

4.2.2 「既知」状態および「未知」状態における行動戦略の効果

我々の提案する行動戦略の効果を確かめるため，「既知」「未知」それぞれの状態において，前述の「ランダム行動」「前進行動」との比較を行った．これらの行動は，まったくのランダム行動に対して「少なくとも移動する」というヒューリスティクスを加えた「ランダム行動」，さらに「後戻りはしない」というヒューリスティクスを加えた「前進行動」という関係になっている．これ以上のヒューリスティクスを環境知識なしで事前設計するのは難しい．既存手法の一つである強化学習については，4.5.2 項で総合的な比較実験を行っている．

さて，これらの戦略ごとに，エージェントが最初に「未知」状態に入ってから学習が収束するまでの実行ステップ数を計測した．同じ状況で 10 回試行し，実行ステップ数の平均と分散を計算して，各戦略の探索能力を評価した．

Fig. 6 に実行ステップ数の平均を示す．この実験では，少ない実行ステップ数のほうが早く学習が収束していることを意味する．

Fig. 6 (a) の結果は，「既知」状態においてオープン端探索戦略が一番効果的に学習を駆動できていることを示している．一方で Fig. 6 (b) の結果によれば，未知状態において同定行動が一番効果的な学習メカニズムであることを示している．

これらの結果から，メタ認知的知覚に基づき行動戦略を切り

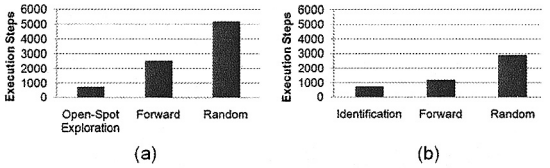


Fig. 6 (a) Efficiency evaluation of open-spot exploration in known region (b) Efficiency evaluation of identification strategy in unknown region

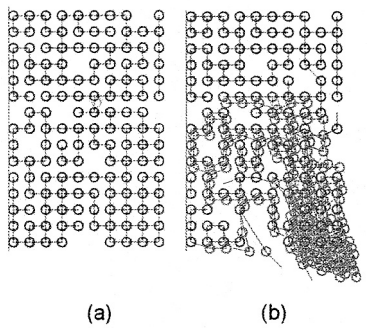


Fig. 7 Effects of learning strategy regulation (a) A result of proposed method. Hidden states are perfectly modeled (b) A result of normal incremental learning using the same data. Many redundant states (drawn in the same position with a slight shift) show the agent's confusion in understanding recent experience

替えながら，不確実性を調整するという提案手法が，環境を探索する上で非常に効果的であると結論づけられる．

4.3 学習戦略のメタ認知的調整の効果の検証

次に，環境モデル学習において，学習戦略の調整の効果を調べた．エージェントが「未知」状態であっても，内部状態を追加して毎回全体学習をかけるエージェントと提案モデルを比較した．基本的な実験条件は前節と同じで，全体学習のタイミングだけを変えた．実験は各戦略ごとに 10 回行った．

Fig. 7 に，オンラインで動きながら追加学習することによって得た内部モデルを示す．提案する全体学習のタイミングでは，700 ステップで新規領域の内部モデルを構築することができた．エージェントは「未知」状態から同定をすることで，「既知」状態に戻ったときの全体学習の結果，冗長な内部状態は少なく，学習の収束も安定していた．

一方で，提案手法と同じ行動を再生しながら，毎回全体学習をするエージェントの結果が Fig. 7 (b) に示されている．この場合，同じ入出力時系列にもかかわらず，同一かどうかを判定できずに融合されていない状態が多数残っている．さらに「既知」領域については，壊れている箇所もでてきており，いつでも毎回学習することで誤解をして元々の知識を壊してしまうという現象も観測できた．

この結果から，学習戦略の自己調整機能は新規領域の学習に対して能力と効果の両方に対して重要な役割を担っていることが分かった．

4.4 非連続な変化がある環境での学習と行動

次に，データに対する仮定をまったく置かずに学習できることを利用して，非連続な変化がある環境での学習を検証した．

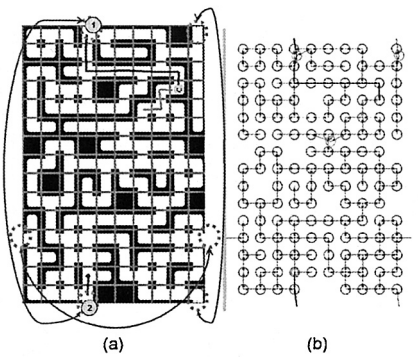


Fig. 8 An environment with discontinuity #1: Torus world

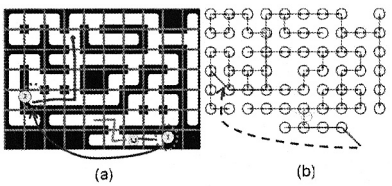


Fig. 9 An environment with discontinuity #2: Warp spot

Fig. 8 (a) で示す上端と下端，右端と左端がつながっている環境（トーラス環境）と **Fig. 9** (a) で示すワープする箇所がある環境（ワープ環境）で検証を行った．

トーラス環境では，上端と下端が 2 箇所，左端と右端が 1 箇所つながっており，上端でさらに上にいく行動を出力すると下端に移動するようになっている．またワープ環境では，右下の点線にある位置から左の真ん中にある点線の位置に一方的に移動する．このような環境は，ロボットのナビゲーション [17] のようにエージェントが行動における上下左右の関係性を前提としていると難しいが，提案手法ではそのような事前知識を仮定しないので，どちらも特別な変更なく学習ができる．Fig. 8 (b), Fig. 9 (b) に学習した内部状態を示すが，どちらも正しい内部構造を獲得していることが分かる．

以上より，非連続な環境であっても，提案手法で内部に環境モデルを構築し，それを利用できることが示せた．これは，迷路の地図だけではなくリモコン操作履歴や抽象度の高い人の行動履歴など非連続な変化をする時系列データの構造化に応用できる可能性があることを示している．

4.5 広い新規領域に対する環境モデル学習

4.5.1 自己調整学習の効果と安定性の検証

Fig. 10 (a) に示す大きな環境で提案手法による自律発達学習を行った．実験環境の大きさ以外はこれまでの実験設定と同じで，左上の 6×10 の小さな領域を初期学習した HMM を内部モデルの初期値として用い，20×30 の領域を学習した．エージェントは，自己調整学習メカニズムを使って自律的に探索したり追加学習したりして，環境を学習する．3,723 ステップ実行したところで，Fig. 10 (b) に示すように環境を完全に学習できた．最終的に学習できた HMM は，587 個の内部状態を持ち，100 万以上のパラメータ（$587 \times (587 \times 5 + 16)$）を正しく設定することができた．

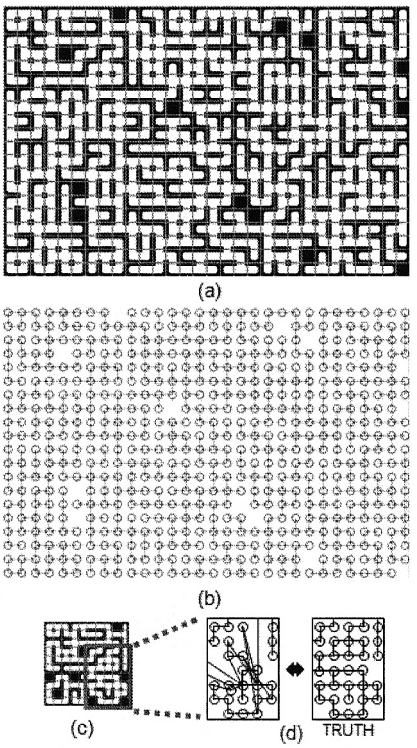
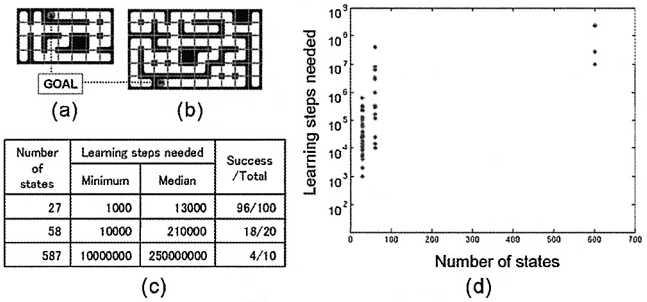


Fig. 10 HMM learning results of significant-scale environments (a) 20 × 30 world used for evaluation of our approach (b) Learning result of our approach: all of many confusing areas are perfectly distinguished in less than 4,000 steps. The agent can move to anywhere using the shortest path (c) 12×10 world used for batch learning (d) Result of batch learning using 20,000-step sequence: confusions of similar areas are not cleared up

次に，提案手法を用いたオンライン発達学習の有効性と比較するために，これまでの実験で用いているのと同じエルゴディック HMM を用いてデータを事前に準備して行うバッチ学習を行った．Fig. 10 (c) に示すそれほど広くない環境において「ランダム行動」と「前進行動」を行い集めた 20,000 ステップのアクション列と観測列を使って全体学習を行ったところ，10 種類のデータ列いずれにおいても，きれいな学習結果は得られなかった．HMM の学習パラメータを調整して何度か試行したが，どれ一つとして成功しなかった．結果の一つを Fig. 10 (d) に示す．'TRUTH' として示すようなきれいな結果が得られるべきであるが，実際には図のように，隠れ状態を正しく見分けることができず，混乱した学習状況にとどまっている．

この差は，特に未知の状況における行動戦略と学習戦略により，安定した追加学習のための質のいいデータ獲得と無駄のない追加学習ができ，学習の収束も早いことから生まれると考えられる．行動面では 4. 2. 1 項で示したように，エントロピーを指標に用いた同定行動算出のおかげで主観的不確定性が変化しないと思われる無駄な行動は選ばれないこと，また「既知」状態で，すばやく新しい領域へと向かうことによる．このような過去の経験をうまく利用した自律発達学習のおかげで，より少ないステップ数で広い領域を学習できることが分かる．



Number of states	Learning steps needed		Success /Total
	Minimum	Median	
27	1000	13000	96/100
58	10000	210000	18/20
587	10000000	250000000	4/10

Fig. 11 Experimental environments and results of reinforcement learning (a) Small world with 27 states (b) Middle-size world with 58 states Other than those shown here, a larger environment shown in Fig. 10 (a) is also used (c) Learning steps needed and success rate for each environment (d) Relation between world size and elapsed time steps for learning (in log scale)

4. 5. 2 強化学習による対照実験

本項では，個々のメカニズムレベルではなく，総合的な性能評価のベンチマークとして強化学習による対照実験の結果を報告する．完全に同等条件での比較は難しかったため，強化学習エージェントの解くべき問題設定は次のように簡略化してある．

- 現在地を特定するセルインデックスが直接観測でき，隠れのない MDP 問題を解けばよい（提案手法は POMDP 問題を解いている）．
- ゴールは特定の 1 箇所に固定し，そこへ向かうような行動さえ獲得できればよい（提案手法では任意の位置から任意の位置へ自由に最適遷移可能な構造を獲得できている）．
- 探索効率化のために，壁方向への移動など，状況を変化させない無駄なアクションは事前に抑止する（提案手法では，そのようなタスク依存の事前知識は利用しない）．

学習には Q 学習法（学習率 0.5，割引率 0.95），行動選択には ε-greedy 法（$\varepsilon = 0.2$）を用いた．エージェントはゴールに到達するごとに報酬 1 を獲得し，ランダムな場所へとジャンプする．実験環境は **Fig. 11** (a), (b) および Fig. 10 (a) に示す 3 種類の環境とし，学習結果の成否は，ランダムに 100 箇所に配置したエージェントがゴールまで到達できた割合が 95%以上か否かで判断した．

Fig. 11 (a) に示す小さな環境ではおよそ 1 万ステップのオーダーで，また (b) の環境では 20 万ステップ程度の時間で構造学習が可能であった．ただし，1 億ステップを超えても正しい構造を獲得できないケースが時々発生しており，(d) に示すとおり，試行ごとのステップ数のばらつきもかなり大きい．提案手法で検討した Fig. 10 (a) の環境では，10 億ステップ経過したあとでも Q テーブル内には何の構造も見られないことが多くあり，この先どの程度時間がかかるものか予想できない．ばらつきの大きさと比べて試行数が十分でないため確定的なことはいえないが，Fig. 10 (c), (d) からは，問題の規模に対して線形以上のオーダーで学習時間が増大し，スケーラビリティに課題のあることがうかがえる．なお，探索効率化のための事前知識を導入しない場合，Fig. 11 (a), (b) のような小さな環境であっても 10 億ステップのうちに学習が成功することは少なく，統計

的に有意なほどの試行を行うことはできなかった．

文献 [18] では，Fig. 10 (a) と近い大きさの POMDP 環境設定で，特定のゴール 1 箇所への到達行動を約 1 億ステップで学習可能なアルゴリズムが報告されている．そのエージェントは「左側に壁が続く限り前進する」「1 行目のインストラクションへ戻る繰り返しループを 5 回実行する」などの，事前に準備された抽象度の高いアクションを利用しているため，問題の複雑さを今回と単純に比較することはできないが，学習効率のスケール感としては前記強化学習の参照実験と大きくは変わらず，我々の得た結果とは依然として差が大きい．

5. 考　　察

4.5.1 項の実験の結果，バッチ学習では困難な規模の広大な環境が，オンライン学習で安定的に学習できることが示された．このことは一見直感に反するようではあるが，我々の提案する自己調整学習の持つ以下の性質を考えると自然である．(1) 行動戦略の切り替えにより，不確実性が適切に制御され，その結果，個々の学習器の状態に応じたデータが供給されることで学習効率が高く維持される，(2) 学習戦略の切り替えにより，新規の経験を安定的に内部モデルに反映できる，(3) 既知状態ではもちろん，未知状態においても過去の経験をうまく利用した行動戦略および学習戦略を実現し，効率的な学習が可能となる．これらの性質により，少ないステップ数で新しい記憶構造を獲得でき，かつオンライン学習での安定性を確保することができるようになった．また，強化学習との比較についても 4.5.2 項で示すように実験を行い，学習に要するステップ数の比較でかなり効率がよくなったことも示すことができた．

ただ，今回の実装手法では，記憶構造の各所に似た状態が独立して存在し，それらをまとめた高次の汎化された構造が明に作られているわけではない．例えば，新しい観測シンボルが来たときの行動生成にこれまでの経験は使えず，ランダムな行動を出力するしかない．しかし，2.3.2 項で示すように，記憶構造の中に点在している似た状況を認識によって取り出し，それらの持つ確率構造を統合することで，現在の状況に似た状態に共通して使える知識を作りだし，未知状態でもランダムではない過去の経験を利用した行動が出力可能になった．

次に，我々のメカニズムがタスクによらず汎用であることを議論したい．自己調整学習の定式化では，アクションや観測シンボルについて何も仮定を置いていない．値も関係性もどのようなものでもよい．さらに，行動やメタ認知などの算出には HMM の内部パラメータだけを用いていることも汎用に使えることを示している．このことから，自己調整学習はアクションと観測の時系列データを持つ様々なタスクに応用可能である．振り子の振りあがりタスクや，物体操作，人とエージェントのインタラクションモデルなど適用範囲は広い．本研究の実証実験では，効果がわかりやすいということで迷路タスクを用いたがこれは 1.1 節で述べたようにより一般性の高いタスクとして設定されている．抽象的な意味においては，エージェントがタスクを遂行するということは何らかの内部構造を使って目的の状態になるように計画し，その計画を実行することであるからである．さらに，4.4 節で示したように非連続な環境も構造化できる．

一方で，提案手法で学習が難しいものとして，構造を持たない時系列データ，例えば偶然に左右される要素の強いものがあげられる．これは，統計的に有意な関係性を持っておらず構造化が難しい．また，現在の実装方法では，新しいアクションを創発することはできない．自分が実行できると知っているアクションから選ぶことはできるが，自分で新しいアクションを創り出し，新しいシンボルを割り当てることはできない．これは連続アクションへの拡張時に同時に取り組む課題である．

ノイズの影響についても簡単に考察する．今回行った実験では，観測にもアクションにもノイズはのせていない．そこで，観測がある一定の割合でまったくランダムになるケースでの予備検証を行った．エージェントが間違った観測を得た場合，メタ認知的知覚は「未知」になるので，新しい内部状態が間違った観測と結びつけられて追加される．ノイズの割合がそれほど高くない場合，例えば 5%ぐらいのときは，オープン端探索やスプリット・マージ法などの学習メカニズムが学習モデルの不必要な冗長性を減らすように働くため，学習やプランニングにほとんど影響はなかった．しかし，ノイズの割合が 10%ぐらいになると，間違った観測によって作り出される内部状態の増加のほうが早くなり，エージェントの状態認識に影響を及ぼし始め，その結果，多くの冗長な状態が未解決のまま残されてしまった．これは現在の提案モデルの短所であり，ノイズのあるデータのモデル化や取り扱いについては改善が必要である．

強化学習との比較は前章で述べたが，既存研究との比較についてもう少し議論を行う．もっと広い意味でのメタ認知の利用という点で谷ら [19] [20] の研究やら門根ら [21] の研究があげられる．両研究とも，エージェントが時系列データから自律的な階層化を行うところにフォーカスが当たっている．階層化は，オープンエンドな学習で大事な課題の一つであり，先に言及したとおり本論文では取り扱っていない範囲である．一方で本論文での主眼は自律的な行動と学習の関係性にある．どちらの研究もデータが与えられた際の構造化に焦点があるが，データをどのように獲得するか，発達の過程自体についての言及はない．人とともに生活する自律発達エージェントでは，オンラインでのエージェントの行動がどれだけ賢く見えるのか，という観点も重要になる．実際に，その場でのエージェントの行動を人が見ることになるので，行動の賢さやその結果として現れる学習に要する試行数が実用上の課題になるからである．我々はその点に特に焦点を当てており，実験中のエージェントは分かりきった場所では時間を使わず，知らないところを次々と開拓していく，というように意味を持って学習が進んでいく実感を得られた．

最後に今後の展望について述べる．自律発達の要件として，(1) 獲得された内部モデルに基づき駆動される自律探索によって学習に有益な時系列データを収集すること，(2) 得られたデータをオンラインで追加学習して，外界の構造を反映した内部モデルとして絶えず更新し続けること，そして (3) 一連の行動および学習を駆動する目的・評価関数を自律的に内部生成することがあげられると我々は考えている．本論文では，(1) (2) を主に取り上げたが，(3) については，1.1 節で紹介した MINDY の技術 [9] と組み合わせて，複数の目的関数を生成し切り替えながら発達していくメカニズムを作っていく．また，複数モー

ダル間の因果関係を学習するという意味での階層化についても研究を進めている．

6. 結　　論

本論文では，オープンエンド環境において自律的に継続的な学習を行う自律発達エージェントのための自己調整学習メカニズムを提案した．

重要なポイントは，現状の理解に関する主観的な不確定性を調整するために行動戦略と学習戦略をそれぞれ切り替える，というメタ認知的な認識と調整である．提案手法の効果は，行動戦略と学習戦略のそれぞれについて，比較実験によって示された．エージェントは，自分の局所的な観測とアクションの時系列情報だけを用いて，隠れ状態を持つある程度大きな環境においても効果的な自律探索と学習を実現できた．さらに，その構造化されたモデルを用いて，未知状態で効果的な同定行動（不確実性を減少させるための行動）を作り出すことができた．

今後の課題としてはこのメカニズムを (a) 連続的なアクションと観測をもつエージェントに適用する，(b) 異なるダイナミクスをもつ環境に適用する，ことがあげられる．自己調整学習メカニズム自体はどちらの課題についてもそのまま適用できると考えているが，(a) については内部モデルの拡張が必要である．特に連続アクションに対応する拡張は今後の課題である．またノイズに対する適応も考慮したい．

上記にあげた拡張を行い，物体操作など違うダイナミクスを持つ問題にも応用することで，オープンエンドな環境において，「未知」状態から「既知」状態へと追加学習をしつつ世界を広げていく，自己発達するエージェントの実現に向けて，知能モデルの探究を続けていく．そして，その探究を通して，人の発達スパイラルの一端を究明していきたいと我々は考えている．

参 考 文 献

[1] M. Lungarella, G. Metta, R. Pfeifer and G. Sandini: “Developmental robotics: a survey,” Connection Science, vol.15, no.4, pp.151–190, 2003.

[2] M. Asada, K. Hosoda, Y. Kuniyoshi, H. Ishiguro, T. Inui, Y. Yoshikawa, M. Ogino and C. Yoshida: “Cognitive developmental robotics: a survey,” IEEE Transactions on Autonomous Mental Development, vol.1, no.1, pp.12–34, 2009.

[3] M. Fujita: “Intelligence dynamics: a concept and preliminary experiments for open-ended learning agents,” Autonomous Agents and Multi-Agent Systems, 2009.

[4] K. Sabe, K. Kawamoto, H. Suzuki, K. Minamino and K. Hidai: “Reward-free Learning using Sparsely-connected Hidden Markov Models and Local Controllers,” The 9th International Conference on Epigenetic Robotics, 2009.

[5] R.S. Sutton and A.G. Barto: Reinforcement Learning: an introduction. MIT Press, 1998.

[6] Ö. Şimşek and A.G. Barto: “An intrinsic reward mechanism for efficient exploration,” Proceedings of the 23rd International Conference on Machine Learning, pp.841–848, 2006.

[7] P.Y. Oudeyer and F. Kaplan: “Intelligent Adaptive Curiosity: a source of Self-Development,” Lund University Cognitive Studies, pp.127–130, 2004.

[8] P.Y. Oudeyer, F. Kaplan, V.V. Hafner and A. Whyte: “The playground experiment: Task-independent development of a curious robot,” Proceedings of the AAAI Spring Symposium on Developmental Robotics, pp.42–47, 2005.

[9] 佐部浩太郎：‘インテリジェンス・モデル MINDY の提案’，身体を持つ知能—脳科学とロボティクスの共進化．pp.197–244，シュプリンガー・ジャパン，2006.

[10] M. Csikszentmihalyi: Flow: The Psychology of Optimal Experience. Harper and Row, 1990.

[11] B.J. Zimmerman: “Self-Regulated Learning and Academic Achievement: An Overview,” Educational Psychologist, vol.25, pp.3–17, 1990.

[12] A. Cassandra, L. Kaelbling and J. Kurien: “Acting under Uncertainty: Discrete Bayesian Models for Mobile-Robot Navigation,” Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems, pp.963–972, 1996.

[13] L.R. Rabiner: “A tutorial on hidden Markov models and selected applications in speech recognition,” Proceedings of the IEEE, pp.257–286, 1989.

[14] L.E. Baum, T. Petrie, G. Soules and N. Weiss: “A Maximization Technique Occurring in the Statistical Analysis of Probabilistic Functions of Markov Chains,” The Annals of Mathematical Statistics, vol.41, no.1, pp.164–171, 1970.

[15] L. Chrisman: “Reinforcement Learning with Perceptual Aliasing: The Perceptual Distinctions Approach,” Proceedings of the Tenth National Conference on Artificial Intelligence, pp.183–188, 1992.

[16] P.R. Cavalin, R. Sabourin, C.Y. Suen and A.S. Britto Jr.: “Evaluation of Incremental Learning Algorithms for an HMM-Based Handwritten Isolated Digits Recognizer,” Proceedings of The 11th International Conference on Frontiers in Handwriting Recognition, pp.1–6, 2008.

[17] S. Thrun, W. Burgard and D. Fox: Probabilistic Robotics (Intelligent Robotics and Autonomous Agents). The MIT Press, 2001.

[18] J. Schmidhuber, J. Zhao and M. Wiering: “Shifting Inductive Bias with Success-Story Algorithm, Adaptive Levin Search, and Incremental Self-Improvement,” Machine Learning, vol.28, pp.105–132, 1997.

[19] J. Tani: “The Dynamical Systems Accounts for Phenomenology of Immanent Time: An Interpretation by Revisiting a Robotics Synthetic Study,” Journal of Consciousness Studies, vol.11, no.9, pp.5–24, 2004.

[20] J. Namikawa and J. Tani: “A model for learning to segment temporal sequences, utilizing a mixture of RNN experts together with adaptive variance,” Neural Networks, vol.21, pp.1466–1475, 2008.

[21] H. Kadone and Y. Nakamura: “Segmentation, Memorization, Recognition and Abstraction of Humanoid Motions base on Correlations and Associative Memory,” 2006 IEEE-RAS International Conference on Humanoid Robots (Humanoids2006), pp.1–6, 2006.

付録 A. 可変ウィンドウ長認識

(1) 初期化: ウィンドウ長 $w = 0$ とし，内部状態に関する事前確率 π_t を一様分布とおく．

(2) 現在の観測を加味して事後確率 $P(Z_t|\pi_t, \boldsymbol{o}_t, \lambda)$ を求め，さらにそのエントロピー $H_w(P(Z_t|\pi_t, \boldsymbol{o}_t, \lambda))$ を計算する．

(3) ウィンドウ長 w が十分長く（$w > \theta_w$），かつ，各ウィンドウ長で計算したエントロピーの系列 $\{H_0, \ldots, H_w\}$ が収束しているなら，最後の認識結果を最終結果として返す．
収束しないままウィンドウ長 w が長くなりすぎたら（$w > \theta_{\mathrm{abort}}$），認識結果不明のまま評価を中止する．
どちらでもなければウィンドウ長 w を 1 増やす．

(4) ウィンドウ長 w まで過去に遡り，シーケンス $\mathcal{U}_w \equiv \{\boldsymbol{u}_{t-w}, \ldots, \boldsymbol{u}_{t-1}\}, \mathcal{O}_w \equiv \{\boldsymbol{o}_{t-w}, \ldots, \boldsymbol{o}_{t-1}\}$ を考慮した事

前確率 $\pi_t = P(Z_t|\pi_{t-w}, \mathcal{U}_w, \mathcal{O}_w, \lambda)$ を計算し直してステップ（2）に戻る．w だけ前の時刻における初期確率 π_{t-w} としては一様分布を用いる．

付録 B. 既知／未知のメタ認知的知覚

(1) 2.1 節で説明した方法を用いて，w ステップ前から現在時刻 t に至るまでの内部状態 Z の時系列 $\{\ldots, z_{t-1}, z_t\}$ を計算する．もし途中で認識結果不明となってしまったのであれば，現状は「未知」状態と判定できる．それ以外の場合は次のステップへ進む．
(2) エントロピー $H(P(Z_t))$ を評価する．もし，$H(P(Z_t)) > \theta_H$ と閾値を超えているのであれば，「未知」状態である．
(3) Z の時系列のもとで，行動観測系列の生起確率を評価する．適切な閾値 $\theta_{P_o}(\lambda), \theta_{P_u}(\lambda)$ のもとで $P(\boldsymbol{o}_t|z_t) < \theta_{P_o}$ または $P(z_t|z_{t-1}, \boldsymbol{u}_{t-1}) < \theta_{P_u}$ となるような時刻 t が存在するなら，「未知」状態である．
(4) 上記のどれにも当てはまらなければ「既知」状態である．

付録 C. 探索／同定戦略の実装

探索／同定戦略は $\arg\{\max_{\mathcal{U}}, \min\} \sum_{\mathcal{O}} P(\mathcal{O}) H(Z|\mathcal{U}, \mathcal{O})$ で表される期待エントロピーの最大／最小化である．これを実現するには以下の手順を用いる．

(1) 先読み深さ $d = 1$ とする．また，探索打ち切りの閾値を θ_H，最大の先読み深さを $d_{\max}$ とする．
(2) d 個のアクション列 $\mathcal{U}_d \equiv \{\boldsymbol{u}_t, \ldots, \boldsymbol{u}_{t+d-1}\}$ のすべての組み合わせそれぞれにおいて，起こりうるすべての観測組み合わせ $\mathcal{O}_d \equiv \{\boldsymbol{o}_{t+1}, \ldots, \boldsymbol{a}_{t+d}\}$ を一つ一つ考え，それらの生起確率 $P(\mathcal{O})$ および，目的関数値 $H(Z|\mathcal{U}, \mathcal{O})$ を求めて期待値 $\sum_{\mathcal{O}} P(\mathcal{O}) H(Z|\mathcal{U}, \mathcal{O})$ を計算する．
(3) 先読み深さ d が最大の先読み深さ $d_{\max}$ に達するか，期待値が探索打ち切りの閾値 θ_H を超えたなら，そこで深読みを停止し，現在の候補の中で最大もしくは最小のシーケンスを返す．閾値を超えるものがなければ，先読み深さ d を 1 増やしてステップ（2）に戻る．

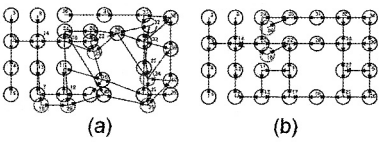


Fig. 12 Effects of "Split-and-Merge" technique (a) A result of normal HMM learning (b) A result of HMM learning with "Split-and-Merge" technique

付録 D. HMM 学習の安定的な収束のためのスプリット・マージ法

スプリット: ある状態において最大確率をもつ観測シンボルの確率値が閾値 p_H（例えば 0.9）以下の場合，その状態を複数に分割し，別の閾値 p_L（例えば 0.1）以上の確率を持つシンボルそれぞれに対して状態を一つずつ割り当てる．分割された各状態への遷移確率はシンボル観測確率に応じて配分され，分割された各状態からの遷移確率は元のものをそのまま引き継ぐ．
マージ: 同じ観測シンボルをもつ複数の状態について，同一のアクション $\boldsymbol{u}$ に対応する遷移が無視できない確率で同一の状態に向かう場合，あるいは同一の状態からのものである場合それらの条件を満たす状態同士を併合して一つの状態とする．

パラメータ推定の過程では，Baum-Welch アルゴリズムの収束後にこの処理を行い，それ以上スプリット・マージ対象の状態がなくなるまで，Baum-Welch アルゴリズムによる最適化とこの処理とを交互に繰り返す．

Fig. 12 に，(a) 一般的な手法による学習結果と (b) スプリット・マージ法を用いた結果を示す．スプリット・マージ法により，正しく区別できなかった状態が正しく分離されたり，冗長だった場所が融合されたりする様子が示されている．

星野由紀子（Yukiko Hoshino）
1998 年東京大学工学系研究科機械情報工学専攻修士課程修了．2001 年同専攻博士課程修了．博士（工学）．同年ソニー株式会社に入社．エンターテインメントロボット QRIO の行動制御アーキテクチャの研究・開発に従事したあと，子供の発達にヒントを得ながら，自律発達エージェントの行動学習，汎化，人とロボットのインタラクションに関する研究に従事．1996 年日本ロボット学会第 11 回研究奨励賞受賞．1999 年同学会第 13 回論文賞受賞．日本赤ちゃん学会正会員．（日本ロボット学会正会員）

野田邦昭（Kuniaki Noda）
2002 年早稲田大学理工学研究科機械工学専攻修士課程修了．同年ソニー株式会社に入社．エンターテインメントロボット QRIO の行動制御モデルの研究開発に従事．2009 年 9 月より 1 年間スイス EPFL に客員研究員として派遣留学．ダイナミカルシステムズアプローチによるロボットの行動学習，スケーラビリティ，汎化，追加学習能力を兼ね備えた学習メカニズム，自律発達する知能の研究などに従事．日本機械学会畠山賞受賞．（日本ロボット学会正会員）

河本献太（Kenta Kawamoto）
1998 年東京大学工学系研究科航空宇宙工学専攻修士課程修了．同年ソニー株式会社に入社．エンターテインメントロボット AIBO，QRIO の画像認識およびシステム設計の研究・開発に従事．その後，人のように柔軟な機械知能の実現を目指し，自律発達学習，因果推定の研究に従事．

佐部浩太郎（Kohtaro Sabe）
1996 年東京大学工学系研究科電気工学専攻修了．同年ソニー株式会社に入社．エンターテインメントロボット AIBO，QRIO の研究・開発に従事した後，自律エージェントの知能の創発に関する研究に従事．専門は画像認識，統計学習，ロボットソフトウェアアーキテクチャなど．2002 年第 8 回画像センシングシンポジウム論文賞受賞．