特　集「Affective Computing」

Deep Emotion：感情理解へ向けた深層感情モデルの開発

Deep Emotion: Development of an Emotion Model Using Deep Learning for Understanding Emotions

日永田　智絵　奈良先端科学技術大学院大学
Chie Hieida　Nara Institute of Science and Technology
hieida@is.naist.jp, https://www.hieida.com/

Keywords: emotion, human-robot interaction, interoception, machine learning, constructive approach.

1. はじめに

人工知能における感情の研究は近年注目を集めている．著者は2019年より人工知能学会全国大会にて「感情とAI」というセッションを企画し，開催してきたが，2年連続で参加者が100名を超え，セッション会場に入りきらないほどであった．解説記事として執筆した「OS-18 感情とAI」[日永田 19]にも，多くの反響があり，感情とAIというコンテンツが非常に注目度が高いことを確信した．

人工物における感情の研究は，人工知能という言葉が多くの人々に広く認知される前から，数多く推進されてきた．代表的なのは2000年頃のPicardによるAffective Computing [Picard 97]やBreazealのkismetという感情ロボットである[Breazeal 02]．日本でも尾形が情動ロボットに関する研究を行っていた[Ogata 00]．当時はaibo [Hornby 00]の発売時期とも重なり，aiboに触発されるように人工物における感情・情動の研究は活発化していく．しかしそれも，日本では2008年頃から停滞する．Google scholarでは2008年頃にそれまで増加傾向だった感情と人工物に関する国内論文が減少する．英文誌ではそのような傾向はないため，日本特有であったのかもしれないが，2006年の第一世代aiboの終焉に合わせるような形で一度ブームが去っているように見える．

それがなぜ今になって注目度が上がっているのだろうか．著者の主観としては，人工知能への注目の高まりと収束にあるように思う．2007年頃から始まる第三次人工知能ブームにより，広く人工知能という言葉が知られ，あらゆる場所で人工知能という言葉を聞く機会が増えた．しかし10年を経て，そのブームも落ち着いてきているように感じる．ある意味ふるいにかけられているのかもしれない．その中で，人工知能の次の一手を期待している人が多いのではないだろうか．オンライン開催となった2020年の人工知能学会全国大会では，「説明性AI」のセッションが非常に多くの聴講を集めたと話題になった．説明性AIも人工知能の次の一手といえるし，参加者が人工知能の行く先を期待し，模索しているように感じる．

人工物は感情をもたない．これは現在非常に一般的な考え方だろう．では，自分自身が感情をもつ，あるいは目の前にいる他者が感情をもつというのはどのように証明できるのだろうか．感情は自分自身の中にある現象のはずだが，その実態は自分自身でもうまく捉えることができない．ひどく曖昧に感じて，自身が「怒っている」と感じているときですら，その理由を捉えることができなかったり，「怒っている」という状態を他者に詳細に説明することは，いくら言葉を尽くしても完全にはできないだろう．こうした強い実感（素朴心理学）が感情についての研究活動の阻害となっている側面がある．

そこで，感情という言葉にとらわれず，少し視点を変え，物事の理解について考える．記号創発ロボティクスでは，経験によって得ることのできるマルチモーダルな情報をカテゴリー分類し，その分類を通した予測によって物事を理解する，つまりこの自己組織的に行われる「マルチモーダルカテゴリゼーション」によって形成された概念が，物事を理解するための基盤として重要な役割を果たすと述べている[長井 12]．このように，理解が個人の経験を通して形成される概念に基づいているとするならば，各自の置かれた環境や身体性によって異なることになり，同一の個体であっても，経験とともに動的に変化する．このように考えると，「理解」が人によって異なることや，ダイナミックに変化すること，一方で理解が身体性に基づいていることが人間どうしの共通の理解を支えていることが納得できる．この考えのもと，長井らは視覚や聴覚，触覚情報といった複数のモダリティの情報を統合し，物体や動作の概念形成に取り組んでいる[長井 12]．

感情もこの物体や動作の概念形成と同様の考え方ができる．感情の心理学的・生理学的知見として，サスペンションブリッジ効果 [Dutton 74] や感情の二要因理論 [Schachter 62] があるが，これらの理論では感情が身体反応だけでなく，そのときの外的状況に影響されるとしている．すなわち，身体の外から得られる五感情報，いわゆる外受容感覚と，臓器などの身体内部の感覚，いわゆる内受容感覚が情報として統合され，感情概念が獲得されると考えられる．つまり，ペットボトルなどの物体も我々が複雑怪奇だと考える感情も，概念形成の観点からいけばそれほど大きな違いはないと考えることができるということである．我々は互いに感情があるということをインタラクションを介して外側から観測している．そして，観測されることによって，共通の概念として形成されていく．したがって，感情を捉えるためには，自己による身体内部の観測だけでなく他者による外側からの観測が重要な要素となる．

そこで，本研究では，外部からの観測の例として，養育者と幼児のインタラクションを題材とし，養育者との社会的なやり取りの中で，感情分化を行うことのできる統合的な感情モデルの構築を目指す．具体的な方法として，既存の統合的な概念モデルをもとに，感情モデルを構築し，それを深層学習で実装する．そのモデルを用いて，養育者とのインタラクションを模したミラーリングタスクを行い，感情分化のシミュレーションを行う．すなわち本研究は，感情メカニズムの理解へ向けた構成的アプローチであると捉えることができる．

神経科学や心理学では，いくつかの統合的な概念モデルが提案されているものの，それらは大人のデータにより，発達後を切り取って見ている．また，工学的なアプローチとして，強化学習を中心とした感情の計算モデルの研究も存在するが [Moerland 18]，これらの研究の多くも感情が発達するということは考えられていない．しかしながら，基本感情にも文化差があることが示唆されているように，感情は初めから出来上がっているとは考えにくく，幼児から徐々に感情が発達すると考えるのが自然である．このような考えの一つとして，Bridges や Lewis などの感情分化の研究があげられる [Bridges 32, Lewis 00]．これらの研究では幼児の振舞いの観察などにより，カテゴリカルな感情が徐々に分化する様子をモデル化している．このようなカテゴリカルな感情は他者から見たときのカテゴリーであり，養育者との社会的なやり取りの中で，概念として固定化していくと考えられる．よって，本研究では，発達の要素を加味して，感情モデルを構築する．

本取組みは，感情メカニズム解明への第一歩に過ぎないが，最終的に感情の仕組みが解明されれば，人間の本質的な理解に近づくことができるだろう．また，本研究の重要な点は，他者とのインタラクションによって感情概念が形成される点にある．他者と共通の概念の形成によって人工物による共感の実現や他者に合わせた感情表出や認識といった有用性が考えられる．従来の手動でつくり込む方法と異なり，つくり手が想定しない行動が Human-Robot Interaction を活性化させる可能性もあるだろう．

2. 感情に関する理論

まずは感情の定義について述べる．感情の定義について普遍的なコンセンサスは著者が知る限りでは存在しないが，本論文では基本的には Damasio の情動（emotion）と感情（feeling）の定義に従う [Damasio 03]．Damasio は，情動を一連の身体的反応，内臓および骨格筋の状態変化，および内的状態の変化として定義し，感情を情動状態の認識として定義している．

次に感情モデルを構成するために，感情にとって必要な要素を考える．上記での定義でもわかるとおり，近年の研究では，感情における身体の重要性を明らかにしている．これは，過去に William James も主張したことであり，感情の末梢起源説と呼ばれている [James 1884]．近年の認知神経科学の研究では，身体内状態の認識である内受容感覚（interoception）が，感情の主観的経験の鍵であるといわれている [Terasawa 13]．感情のカルテット理論では，脳幹中心のシステムは感情システムに対応しているとされる [Koelsch 15]．脳幹は最も古い脳構造であり，網様体は脳幹中心のシステムにおいて重要な役割を果たす．感情と身体の関係として重要なもう一つの側面は，Damasio のソマティックマーカー仮説であり，感情が私達自身の身体を通して効率的に外部刺激を評価すると仮定したものである [Damasio 96]．これは，著者が身体内部の評価（内的評価）と外部刺激への評価（外的評価）の両方を同時に検討する動機となった．

上記のとおり，身体は感情の起源であり，不可欠であるといえる．本研究では，身体と内受容感覚を一つのサブシステム，すなわち提案モデルの第 1 層とする．実際，感情の分化特性がありながらも，怒り，喜び，嫌悪感，恐怖，悲しみ，驚きなどの基本的な感情は文化に関係なく存在するといわれる [Ekman 71]．これは，我々人間が類似の身体と環境を共有しているために可能であるとすれば説明がつくし，感情が我々の身体状態に基づいているという事実を裏付けている．また，能動的推論（active inference）のアイディアは，外的刺激の評価に関係する visual attention など，身体システムにも関連している [Friston 10, Seth 16]．

感情が意思決定と原因因果関係の推論に関連しているということも重要な要素だ [Ledoux 98]．例えば，サスペンションブリッジ効果として知られている身体反応の誤帰属は，誰かに会っているときに身体が危険にさらされる恐怖体験をすることによって，誤って自身の身体反応を他者に帰属してしまうというものである [Dutton 74]．

このより高いレベルの認知プロセスは，眼窩前頭野中心のシステムと深く関係していると考えられ，皮質–基底核ループに由来する強化学習モジュールも関連している．能動的推論と強化学習との関係は [Friston 09] で議論されている．本研究では，意思決定を一つのサブシステムとみなし，提案するモデルで第 3 層と呼ぶ．

また，記憶に基づくシステムは，感情モデルの構成要素として重要だと考えられる．カルテット理論でいうと海馬中心のシステムが海馬と扁桃体という主に関与する記憶に基づくシステムに対応している [Koelsch 15]．感情における扁桃体の活動は特に重要であり，長い間研究されてきた．Yakovlev 回路は有名な辺縁系回路の一つであり，扁桃体が回路に含まれている [Yakovlev 48]．Papez 回路も同様に有名な海馬を含む辺縁系の回路である [Papez 37]．これらは回路として独立しているが，皮質，基底核，間脳を介して互いに相互作用をもち，密接に関連している [Mendoza 07]．本論文では，記憶に基づくシステムを一つのサブシステムと考え，提案するモデルの第 2 層と呼ぶ．このサブシステムは，システム全体を環境に適応させるために，生得的な評価システム，すなわち第 1 層に柔軟性を与える．これまでの議論は，感情システムが三つのシステム，1）身体に基づいた評価システム，2）記憶に基づくシステム，3）意思決定に関連する学習システムに分けられていることを意味する．

以上のように感情は，各モジュールが相互に関係しあっているため，局在的な考え方ではなく，ネットワークとして考える必要がある．したがって，上述のサブシステムが統合されていることが重要である．本研究では統合に関して Damasio の概念モデル [Damasio 96] と守口と小巻の心理学的構成主義に基づいたモデルを参考とする [Moriguchi 13, 日永田 19]．ダマシオの概念モデルでは，外受容感覚や内受容感覚を受け，扁桃体がその情報を評価，視床下部らが身体反応を誘発し，実際に身体反応が起こる．これが情動状態である．その情報が内受容感覚として大脳皮質に送られ，外受容感覚と統合され，感情が知覚される．守口と小巻のモデルでは，身体状態から内受容感覚を経て，コアアフェクト（情動状態）が形成される．そのコアアフェクトと文脈や概念といった情報が統合され，カテゴライズされることによって感情が知覚される．このモデルは，アレキシサイミアのニューロイメージング研究の調査に基づき提案されたモデルである．本研究で提案するモデルはこれらのモデルに深く根付いている．しかし，冒頭でも述べたとおり，統合的な概念モデルは，発達後を切り取ってみている．基本感情にも文化差があることが示唆されているように，感情が初めから出来上がっているとは考えにくく，幼児から徐々に感情が発達すると考えるのが自然である．そこで，本研究では，発達の過程として，感情分化に着目し，モデルの構築を行った．次の項にて提案する感情モデルを説明する．

3. 感 情 モ デ ル

前章までの内容に基づき，感情モデルにおける要件は下記のとおりとなる．

- 感情において身体が重要で，身体情報は意思決定を効率化する．
- 感情システムは三つのシステム，1）身体に基づいた評価システム，2）記憶に基づくシステム，3）意思決定に関連する学習システムに分けられている．

また，後述するが，感情モデルの学習における重要な点として，

- 報酬は身体反応によって規定される

と考える．これはホメオスタシスやアロスタシスの考えに基づくものである [Sterling 11]．統合的なモデルの構成は，既存の統合的な概念モデルに従う．まず，守口と小巻のモデルをベースとし，入力として内外からの刺激と，出力として行動を想定する．Damasio のモデルより，入力の刺激が評価され，身体反応を誘発する部分が，守口と小巻のモデルでの身体状態につながる．そして，守口と小巻のモデルで示される，文脈，概念，コアアフェクトがカテゴリー化され，感情として知覚される中で，行動が出力される．これは身体反応が意思決定を効率化するという考えからである．

具体的には，本感情モデルは 3 層に分かれており，刺激に対しすばやく身体反応を返す第 1 層，記憶にアクセスし経験に基づいて刺激を評価する第 2 層，未来予測および行動を出力する第 3 層で構成されている．本論文では実装についての詳細を説明しないため，[Hieida 18] を参照されたい．

第 1 層は身体を使って迅速に刺激に反応する．これを外部評価（external appraisal）と呼ぶ．さらに，第 1 層は外部知覚は関係ない自身の身体状況，すなわち内部評価（internal appraisal）を反映する．この層が感情が身体に依存する理由となる．反応は生得的にあらかじめプログラムされているので，刺激に対して過度の反応を引き起こすことがある．そこで，第 2 層は記憶にアクセスして，刺激が経験を通して評価されるようにする．第 2 層は，不必要な反応を抑制し，同時に重要な問題に迅速に反応することを可能にする．もちろんこれは処理コストと刺激に対する応答の精度との間のトレードオフである．第 2 層によって精度を増した第 1 層の出力は，外界および内的状態の評価結果の次元圧縮したものの知覚として考えられる．この第 1 層の出力の知覚を，内受容感覚（interoception）とみなす．ここでの主な実装は Recurrent Attention Model（RAM）[Mnih 14] という時系列モデルでのラベル付き画像データセットからの教師あり回帰学習である．刺激に対して事前に人手で付けられた感情値を出力するようなモデルを構築し，それを次元圧縮された生得的な反応として置き換える．当然こ

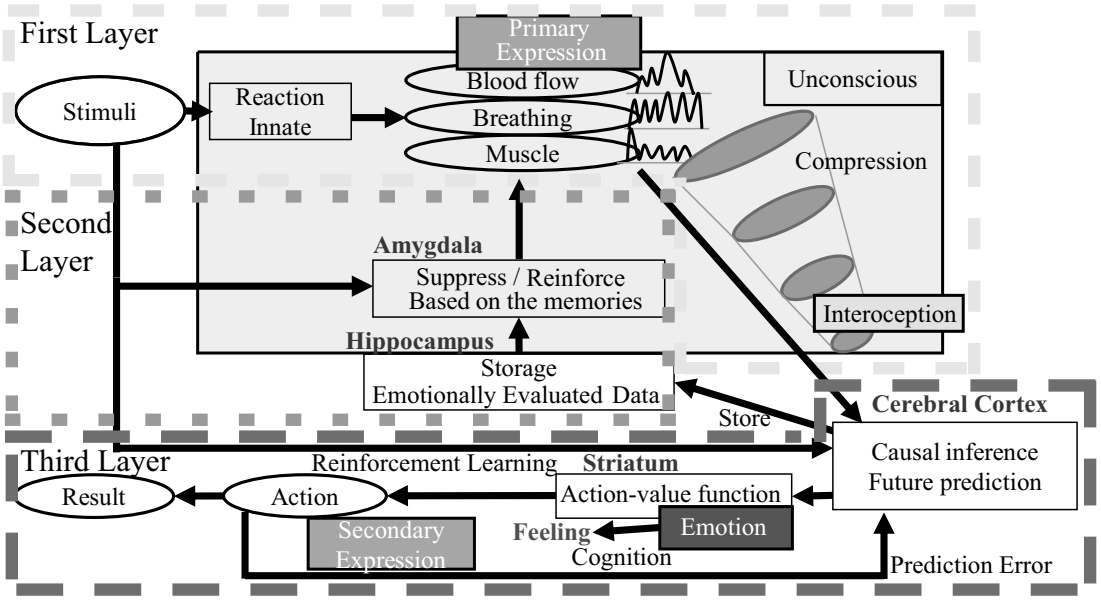

図 1　提案する感情モデル

れはいくつかの問題をはらんでいるが，それは後述の議論で話をする．第 2 層はルールベースによる実装となっており，過去のデータをメモリにため，時系列的に一定以上の差がある場合にそれを補正するように働く．いわば平滑化モジュールといえる．

図 1 に示すように，第 3 層では，第 1 層・第 2 層の出力と刺激を用いて原因推論と予測を行い，入力刺激と予測結果を用いて行動決定を行う．第 3 層の最も重要な部分は，最適な行動決定の学習をする強化学習である．強化学習では報酬の設定が重要となってくるが，提案する感情のモデルでは，内部状態の調節メカニズムである「ホメオスタシス（homeostasis）」の考え方を採用する．これは，モチベーションの基本理論であるドライブリダクション理論に基づいている [Myers 10]．また興味深いことに，ホメオスタシスは，感情のカルテット理論の感情システムの一つである間脳に密接に関連している [Koelsch 15]．つまり，第 1 層の出力，すなわち内受容感覚が一定であるときに報酬が提供される．この一定の条件として与えられる値は永続的に同じ値をとるわけではなく，ある一定期間内での内受容感覚の値の平均を用いる．言い換えるとホメオスタシスは内受容感覚を完全に一定に保つのではなく，急な変化を阻止するように設定される．本研究のモデルでは，時間の経過とともに徐々に変化する一定の条件として与えられる値を「homeostatic setpoint（setpoint）[Keramati 14]」と定義する．これはアロスタシスの考え方に根付いている [Sterling 11]．第 3 層の実装には，Convolutional LSTM-DDPG を用いる．これは時系列モデルである Convolutional LSTM [Xingjian 15] と強化学習手法の一つである Deep Deterministic Policy Gradient（DDPG）[Lillicrap 15] を組み合わせたものであり，予測部分を LSTM が行動出力部分を DDPG が担っている．

行動決定プロセスの後，予測誤差が計算され，第 3 層が更新される．経験はエピソード記憶として海馬に保存され，第 2 層で感情評価が行われる．第 1 層は生得的なネットワークであり，学習プロセスは第 2 層と第 3 層にのみ存在する．守口と小巻のモデルの感情の捉え方に従えば，内受容感覚（身体反応）や予測の情報（文脈）が入力され，行動（概念）によってカテゴライズされる線条体における Policy（方策）の神経パターンが，感情（emotional feelings）として認知されると考えられる [Moriguchi 13]．

重要な点として，本モデルでは，内的および外的評価，すなわち感情の素である身体反応，肉体がなければ，感情はないはずだと考えている．次に，提案したモデルでは予測が不可欠である．そして，ソマティックマーカー仮説に関連する行動決定も重要な要素である．これらの視点は，近年提案された EPIC モデルと非常に近い考え方である [Barrett 15]．EPIC モデルの考え方は，予測符号化と能動的推論 [Friston 07] に基づいている．著者は EPIC モデルとは独立して提案モデルを開発したが，いくつかの重要なアイディアがこれらのモデルで共有されている．本研究で提案された EPIC モデルとモデルの主な違いは，いくつかの深層学習モジュールを組み合わせて，提案モデルの実際の実装を提案することである．逆に，EPIC モデルは概念モデルであり，予測符号化に根付いている．

感情モデルが考慮すべきこととして，Cañamero が主張する人工的な感情システムの設計がある [Cañamero 05]．Cañamero は感情はロボットの身体と社会適応に関係した内部価値体系に基づいていなければならないと主張した．また，モデルは感情，動機付け，行動，知覚，および「認知」のさまざまな側面の間のリンクを設計すべきであり，リンクはエージェントの身体に根ざしてい

なければならない．すでに述べたように，提案した感情モデルはこれらの要件を満たす可能性を秘めている．

4. 実　　験

導入でも述べたとおり，本研究では，カテゴリカルな感情が養育者とのインタラクションの中で形づくられていくということを想定しており，本モデルでターゲットとするのも，どのように感情分化が起きていくかのシミュレーションをすることである．さらにはそれがモデルのどの部分で起きるのかやどのような刺激を入れることによって起こるのかというのも非常に興味深い点である．下記では，実験について説明する．

4·1　エージェントシステム

本実験では，身体として仮想エージェントを用いる．仮想エージェントには，三次元エージェントのモデリングフリーソフトウェアパッケージの“MakeHuman”を使用した（http://www.makehumancommunity.org/wiki/MakeHuman_resources）．MakeHuman は表情に関する制御パラメータを 60 個もつが，本研究では 19 個を使用し，目・眉・口・口角を制御した．視線の変更は第 1 層の実装である RAM の画像に対する attention の値を用いて制御した．その他の口の開閉，眉の上下，口角の上下，瞼の開閉に関しては，第 3 層の DDPG により出力される action を用いて制御した．action は口の開閉，眉の上下，口角の上下，瞼の開閉のそれぞれに対して 0 から 1 の値をとり，それぞれの値は連続的にエージェントの表情を動かす．エージェントのデザインは著者が子供を想定し作成した．

4·2　タスク設計

提案モデルを上記のエージェントに実装し，養育者とのインタラクションのシナリオに基づいて“facial expression”タスクを設計した．このタスクでは，インタラクションパートナー（母エージェント：提案モデル未実装）が，幼児エージェント（提案モデル実装）の表情を四つのカテゴリーの一つとして認識し，同じカテゴリーの表情を幼児エージェントに対して行う．母エージェントによる幼児エージェントの表情認識は，1）喜び（口角が上がるとき），2）怒り（口角が落ちる，眉をしかめた，目が半分以上開いている），3）悲しみ（口角が落ちる，眉をしかめる，目が半分以上閉じている），4）ニュートラル（それ以外）とした．

この実験デザインは，「ミラーリング」と呼ばれる既知の現象に基づいており，この現象では母親は日常的に乳児の表現を直感的に模倣するといわれる [Winnicott 60]．これは，乳幼児が感情調整と社会的反応の学習に重要である [Murray 16]．特に笑顔のために，乳幼児とその保護者との間のインタラクティブスマイルゲームは，乳幼児の社会性発達における重要なマイルストーンとして知られており，後の社会的能力の基盤を構築する [Kaye 80]．また，Ruvolo らは，子供が笑顔になるタイミングと母親との関係について戦略が存在することを明らかにした [Ruvolo 15]．したがって，本実験の目的は，幼児エージェントが学習した行動と，表情戦略の学習におけるモデルの内部状態の変化を観察することである．

本実験では，「顔画像条件（face-only）」と「顔画像＋ノイズ条件（face＋natural）」という二つの異なる条件を行った．これらの条件は，母親の顔だけを見る理想的な状態と，環境要因が存在する場合とを比較するために設定した．「顔画像条件」状態では，幼児エージェントは常に自身の表現（ミラーリング）に従って母エージェントから顔画像の入力を受ける．顔画像は JAFFE データベース [Dailey 10] から選択された 8 枚の顔画像を用いる．各感情カテゴリーには二つの異なる顔画像があり，二つの画像のうちの一つがランダムに選択され，幼児に提示される．一方「顔画像＋ノイズ条件」では，幼児エージェントは，顔画像の一つまたは IAPS 画像 [Lang 99] から選ばれた 8 枚のうちの一つをランダムに画像刺激として入力される．IAPS の画像は環境刺激をイメージしている．顔画像は，幼児エージェントの行動に従って選択されるので，幼児エージェントが操作することができる刺激である．しかし，IAPS 画像は，ランダムに選択されるので，エージェントには直接操作できない刺激である．言い換えると，幼児エージェントは，与えられた顔画像に基づいて，自身が意図した刺激を取得するための方策を学習し，IAPS 画像から望ましくない刺激が提示されたときに目を閉じるなどの対策を学習することが期待される．どちらの条件も，エージェントが目を閉じると，閉じた目の部分の画像が黒い画像として表示される．本稿では「顔画像＋ノイズ条件（face＋natural）」結果のみを紹介する．

提案した感情モデルと上記のシナリオを用いて，100 000 epoch 学習を行った．学習が進むたびに，養育者とのインタラクションを通して幼児エージェントによって構築された状態空間を観察するために，主成分分析（PCA）を用いて第 3 層のポリシーネットワークの中間層を可視化した．著者の感情に対する仮説が正しく，適切に実装されていれば，この状態空間は行動によって感情カテゴリーに分かれることが期待される．

4·3　結　　果

図 2（a），（b）は，内受容感覚（interoception）および DDPG のポリシーネットワークの中間層を PCA で次元圧縮し可視化したものを示している．これらの結果は図 1 の第 1 層の最終出力と第 3 層の行動出力のモジュールに対応する．各色は，母エージェントが認識する表情のカテゴリーを表している．これらの図から，ポリシーネットワークにおける表現は，interoception と比較して，

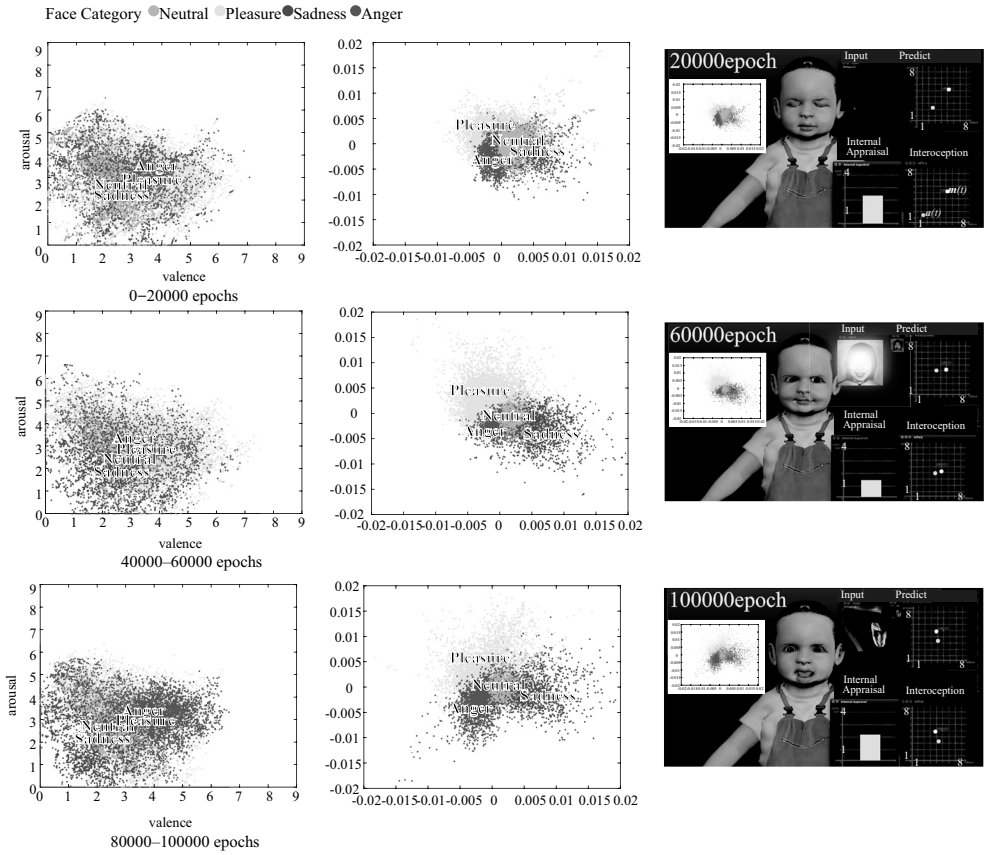

図2 モデルの内部状態.
（a）内受容感覚，（b）ポリシーネットワークの中間層のPCAによる可視化，（c）表情例

感情カテゴリーを非常にはっきりと分けていることがわかる.

この結果を定量的に評価するため，感情カテゴリーの分離度を多変量正規分布を用いて算出した．結果を図3に示す．分離度は各空間に対して下記のように計算した．まず，それぞれの感情カテゴリー（ニュートラル，喜び，悲しみ，怒り）を多変量正規分布と仮定し，分布を計算した．次に，正解カテゴリーが最も高い確率と算出された割合の値の平均を分離度として計算した．したがって，分離度が高いということは空間内の各カテゴリーの重なりが少ないということになり，より重なりが多いと分離度は低くなる．図3より，ポリシーネットワークの中間層の結果は，徐々に分離度が上がっていく様子が観測でき，ポリシーネットワークの中間層のほうがより分離度が高いことがわかる．分離順はまず，喜びとそれ以外で分離し，その後怒りと悲しみが分離しているように見える．これはBridgesの感情分化で述べられている，快と不快が分離し，その後不快が詳細化するという内容に一致している[Bridges 32]．また，ポリシーネットワークの中間層の結果は，ラッセル円環モデルの構造にも類似しており，人間に近い感情構造が見られていると考えることができる．すなわち，ここを感情状態（Emotional state）と置き換えられるのではないかと考察できる．また，記述外ではあるが，興味深いことに，「顔画像条件」よりも「顔画像＋ノイズ条件」のほうが分離度が高くなっていた.

続いて，学習モデルを使用してエージェントの行動を観察した．図2（c）は，特定のepochでの各モデルの幼児エージェントによる典型的な表情を示している．幼児エージェントの行動の観察から，以下のような行動変化があることがわかった*1.

- 20 000 epochs：エージェントはよく目を閉じている.
- 60 000 epochs：エージェントはしばしば目を開き，刺激によって表情を変える.
- 100 000 epochs：エージェントはさまざまな表情（驚き，怒りなど）を出力し，interoceptionを安定させることに成功した.

基本的にエージェントは非常に頻繁に笑顔を浮かべ，相手を笑わせる．この行動はエージェントがパートナーのために，パートナーを笑顔にしようとしているように見える．この現象は[Ruvolo 15]の所見と一致している．実際には，幼児エージェントがただ単に欲求の刺激のために笑っている，つまりエージェントが選択的微笑を学

*1 https://youtu.be/Phjn58kJ2nsにて，エージェントのデモビデオを公開している.

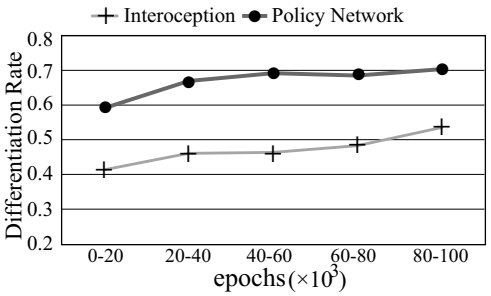

図 3　感情モデルの内部状態のカテゴリー分離度

んだということになり，非常に興味深い結果である．

5. 議　　論

実験では，提案モデルを幼児エージェントに実装し，母エージェントとのインタラクションの様子を観察した．記載外であるが，顔画像条件はノイズ画像が含まれる条件に比べて，分離度が低かった．また，顔画像条件は喜びが半分を占め，残りの半分はニュートラルと悲しみの組合せから構成されているように見え，怒りはより少なかった．しかし，顔画像＋ノイズ条件の結果では，喜びが大部分ではあるが，顔画像条件に比べて怒りが増し，ニュートラルと悲しみが分離される．これは，エージェントが常に制御できる環境では，怒りを出力する必要はないと考えられ，制御不能な刺激のために刺激をより詳細に分類することによって行動を選択する必要があったと考えられる．したがって，制御可能な刺激だけでなく，制御不能な刺激も人間のような豊かな感情をつくり出すと推測される．制御不能な刺激はまた，未来を予測するための学習に非常に重要な意味を与える．つまり，世界が完全に予測するのに十分単純であれば，学習はほとんど意味をもたない．

さらに，図 2（c）の 100 000 epoch の結果は，幼児エージェントがヘビに驚いているように見え，これ自体も非常に興味深い．ポリシーネットワークの中間層の PCA 空間，すなわち感情状態の内部表現において，驚きカテゴリーが生成されたかどうかは，アクションが四つの感情カテゴリーのみで分類されたため，明確ではない．しかし，提案された感情モデルの内部表現として，より豊かな感情空間が出現する可能性がある．この点にはまださらなる分析が必要である．

次に，提案モデルの限界について述べる．第 1 層の学習で用いている画像データベースの IAPS は成人の被験者が valence および arousal のラベルを付けているため，文脈などの事前知識が加味されてしまうという問題が存在する．しかし，今回はラベル値の平均化処理によってデータの個性が低下し，生得的な反応に近い自然反応が抽出することができると考え，実際に過去の実験でも，学習したモデルが幼児と類似する反応をいくつか示している．この手法の別の方向性として IAPS データベースの代わりに実際の人体からの生体信号を RAM のトレーニングに使用するということが考えられる．

また，当然本モデルのようなモデルをどのように評価すべきかという問題もあり，近年の人工知能学会全国大会でも開催しているセッション「感情と AI」[日永田 19] の中での議論では，「感情がつくれた」ということは不可能ではないかという議論もある．これは，我々人間が，他者に面したときに，相手に本当に感情があるといえるかということと同様である．すなわち，感情が関連すると思われる現象のシミュレーションやインタラクションの中での有用性でしか，現状の感情モデルの評価というのは難しいのだろう．そのうえで，モデルの内部の挙動から，神経科学的な仮説が導かれることが望ましい．こうした規模の大きさから，感情のモデル研究には，その研究自体をどのように評価すべきかという問題が常につきまとっているように感じる．可能であれば，コミュニティとして評価指標を確立できれば，この分野はより広がりをもてるだろう．こうした評価とそのタスクの問題は，感情を用いた強化学習エージェントに関する調査論文でも述べられている [Moerland 18].

6. 結　　論

本研究では，養育者との社会的なやり取りの中で，感情分化を行うことのできる統合的な感情モデルの構築を目指した．そして，既存の統合的な概念モデルをもとに，感情モデルを構築し，それを深層学習で実装した．そのモデルを用いて，養育者とのインタラクションを模したミラーリングタスクを行い，感情分化のシミュレーションを行った．結果として，実装した感情モデルのポリシーネットワーク 内に，感情分化と類似した現象が見られ，ラッセル円環モデルと近い構造をもつことを確認した．

今後の課題として，より複雑なタスクを用いて提案モデルの評価を行う．また，人とのインタラクションの中でどのような感情を獲得するか，逆に人に対してどのような影響を与えるのかを検証する．本モデルの応用としては，養育者とのインタラクションにおいて，どのような刺激を与えると，どのような子供の振舞いが学習されるのかといったシミュレーションをすることが考えられる．これは，保育現場などに有用かもしれない．最終的には基本感情だけでなく，複雑な社会的感情の実現できるモデルの構築を目指す．

◇ 参 考 文 献 ◇

[Barrett 15] Barrett, F. L. and Simmons, W. K.: Interoceptive predictions in the brain, *Nature Reviews Neuroscience,* Vol. 16, No. 7, pp. 419-429（2015）

[Breazeal 02] Breazeal, C.: *Designing Sociable Robots,* The MIT Press（2002）

[Bridges 32] Bridges, K. M. B.: Emotional development in early infancy, *Child Development,* pp. 324-341（1932）

[Cañamero 05] Cañamero, L. and Gaussier, P.: Emotion understanding: Robots as tools and models, *Emotional Development,* pp. 235- 258（2005）
[Dailey 10] Dailey, M. N., Joyce, C., Lyons, M. J., Kamachi, M., Ishi, H., Gyoba, J. and Cottrell, G. W.: Evidence and a computational explanation of cultural differences in facial expression recognition, *Emotion,* Vol. 10, No. 6, pp. 874-893（2010）
[Damasio 96] Damasio, A. R., Everitt, B. J. and Bishop, D.: The somatic marker hypothesis and the possible functions of the prefrontal cortex [and Discussion], *Philosophical Trans. of the Royal Society B, Biological Sciences,* Vol. 351, No. 1346, pp. 1413- 1420（1996）
[Damasio 03] Damasio, A.: *Looking for Spinoza: Joy, Sorrow, and the Feeling Brain,* Harvest Books, Harcourt（2003）
[Dutton 74] Dutton, D. G.: Some evidence for heightened sexual attraction under conditions of high anxiety, *J. of Personality and Social Psychology,* Vol. 30, No. 4, pp. 510-517（1974）
[Ekman 71] Ekman, P. and Wallace, F. V.: Constants across cultures in the face and emotion, *J. of Personality and Social Psychology,* Vol. 17, No. 2, pp. 124-129（1971）
[Friston 07] Friston, K. J. and Stephan, K. E.: Free-energy and the brain, *Synthese,* Vol. 159, No. 3, pp. 417-458（2007）
[Friston 09] Friston, K. J., Daunizeau, J. and Kiebel, S. J.: Reinforcement learning or active inference?, *PLoS One,* Vol. 4, No. 7, pp. 1-13（2009）
[Friston 10] Friston, K. J., Daunizeau, J., Kilner, J. and Kiebel, S. J.: Action and behavior: A free-energy formulation, *Biological Cybernetics,* Vol. 102, No. 3, pp. 227-260（2010）
[Hieida 18] Hieida, C., Horii, T. and Nagai, T.: Deep emotion: A computational model of emotion using deep neural networks, *arXiv preprint arXiv:1808.08447*（2018）
[日永田 19] 日永田智絵，堀井隆斗，長井隆行 ほか：OS-18 感情とAI，人工知能，Vol. 34, No. 6, pp. 881-887（2019）
[Hornby 00] Hornby, G. S., Takamura, S., Yokono, J., Hanagata, O., Yamamoto, T. and Fujita, M.: Evolving robust gaits with AIBO, *Proc. 2000 iCRA, Millennium Conference, IEEE Int. Conf. on Robotics and Automation, Symposia Proceedings*（cat. No. 00CH37065）, Vol. 3, pp. 3040-3045, IEEE（2000）
[James 1884] James,W.: What is an emotion?, *Mind,* Vol. os-IX, No. 34, pp. 188-205（1884）
[Kaye 80] Kaye, K. and Fogel, A.: The temporal structure of face-to-face communication between mothers and infants, *Developmental Psychology,* Vol. 16, pp. 454- 464（1980）
[Keramati 14] Keramati, M. and Gutkin, B.: Homeostatic reinforcement learning for integrating reward collection and physiological stability, *Elife,* Vol. 3, p. e04811（2014）
[Koelsch 15] Koelsch, S., Jacobs, A. M., Menninghaus, W., Liebal, K., Klann-Delius, G., Scheve, von C. and Gebauer, G.: The quartet theory of human emotions: An integrative and neurofunctional model, *Physics of Life Reviews,* Vol. 13, pp. 1-27（2015）
[Lang 99] Lang, P. J., Bradley, M. M. and Cuthbert, B. N.: International Affective Picture System（IAPS）: Technical manual and affective ratings, Gainesville, FL: The Center for Research in Psychophysiology, University of Florida（1999）
[Ledoux 98] Ledoux, J. E.: *The Emotional Brain: The Mysterious Underpinnings of Emotional Life,* Simon & Schuster（1998）
[Lewis 00] Lewis, M.: Self-conscious emotions, *Handbook of Emotions,* p. 742, Guilford Press（2000）
[Lillicrap 15] Lillicrap, T. P., Hunt, J. J., Pritzel, A., Heess, N., Erez, T., Tassa, Y., Silver, D. and Wierstra, D.: Continuous control with deep reinforcement learning, *arXiv preprint arXiv:1509.02971*（2015）
[Mendoza 07] Mendoza, J. and Foundas, A.: *Clinical Neuroanatomy: A Neurobehavioral Approach,* Springer New York（2007）
[Mnih 14] Mnih, V., Heess, N., Graves, A. and Kavukcuoglu, K.: Recurrent models of visual attention, *Int. Conf. on Neural Information Processing Systems*（2014）
[Moerland 18] Moerland, T. M., Broekens, J. and Jonker, C. M.: Emotion in reinforcement learning agents and robots: A survey, *Machine Learning,* Vol. 107, No. 2, pp. 443-480（2018）
[Moriguchi 13] Moriguchi, Y. and Komaki, G.: Neuroimaging studies of alexithymia: Physical, affective, and social perspectives, *BioPsycho Social Medicine,* Vol. 7, No. 1, p. 8（2013）
[Murray 16] Murray, L., De Pascalis, L., Bozicevic, L., Hawkins, L., Sclafani, V. and Ferrari, P. F.: The functional architecture of motherinfant communication, and the development of infant social expressiveness in the first two months, *Scientific Reports,* Vol. 6:39019, pp. 1-9（2016）
[Myers 10] Myers, D.: *Psychology,* Worth Publishers（2010）
[長井 12] 長井隆行, 中村友昭：マルチモーダルカテゴリゼーション：経験を通して概念を形成し言葉の意味を理解するロボットの実現に向けて（特集：記号創発ロボティクス），人工知能，Vol. 27, No. 6, pp. 555-562（2012）
[Ogata 00] Ogata, T. and Sugano, S.: Emotional communication between humans and the autonomous robot wamoeba-2（waseda amoeba）which has the emotion model, *JSME Int. J. Series C Mechanical Systems, Machine Elements and Manufacturing,* Vol. 43, No. 3, pp. 568-574（2000）
[Papez 37] Papez, J.: A proposed mechanism of emotion, *Arch Neurol Psychiatry,* Vol. 79, pp. 217-224（1937）
[Picard 97] Picard, R.: *Affective Computing,* MIT Press, Cambridge（1997）
[Ruvolo 15] Ruvolo, P., Messinger, D. and Movellan, J.: Infants time their smiles to make their moms smile, *PLoS One,* Vol. 10, No. 9, pp. 1-10（2015）
[Schachter 62] Schachter, S. and Singer, J.: Cognitive, social, and physiological determinants of emotional state, *Psychological Review,* Vol. 69, No. 5, pp. 379-399（1962）
[Seth 16] Seth, A. K. and Friston, K. J.: Active interoceptive inference and the emotional brain, *Philosophical Trans. of the Royal Society of London B: Biological Sciences,* Vol. 371, No. 1708（2016）
[Sterling 11] Sterling, P.: Allostasis: A model of predictive regulation, *Physiology and Behavior,* Vol. 106, pp. 5-15（2011）
[Terasawa 13] Terasawa, Y., Fukushima, H. and Umeda, S.: How does interoceptive awareness interact with the subjective experience of emotion? An fMRI study, *Human Brain Mapping,* Vol. 34, No. 3, pp. 598-612（2013）
[Winnicott 60] Winnicott, D.: The theory of the parent-infant relationship, *Int. J. of Psychoanalysis,* Vol. 41, pp. 585- 595（1960）
[Xingjian 15] Xingjian, S., Chen, Z., Wang, H., Yeung, D.-Y., Wong,W.-K., and Woo,W.-c.: Convolutional LSTM network: A machine learning approach for precipitation nowcasting, *Advances in Neural Information Processing Systems,* pp. 802-810（2015）
[Yakovlev 48] Yakovlev, P.: Motility, behavior and the brain, Stereodynamic organization and neural co-ordinates of behavior, *J. Nerv. Ment. Dis.,* Vol. 107, pp. 313-335（1948）

2020年11月5日　受理

著 者 紹 介

日永田　智絵（正会員）

2014年電気通信大学情報理工学部知能機械工学科卒業．2016年同大学院情報システム学研究科情報メディアシステム学専攻修士課程修了，修士（工学）．同年，日本学術振興会特別研究員（DC1），2019年電気通信大学大学院情報理工学研究科機械知能システム学専攻博士後期課程単位取得済退学．博士（工学）．同年，大阪大学先導的学際研究機構付属共生知能システム研究センター特任研究員を経て，現在，奈良先端科学技術大学院大学先端科学技術研究科情報科学領域助教．HRI，情動，感情モデルの研究に従事．