

Symposium : On Emotion How to Make a Feeling in Neural Networks

メタデータ	言語: jpn 出版者: 公開日: 2017-10-02 キーワード (Ja): キーワード (En): 作成者: メールアドレス: 所属:
URL	http://hdl.handle.net/2297/11054

シンポジウム「情念について」

感情の作り方（ニューラルネットワークの場合）

柴田 正良（金沢大学）

月本 洋（東京電機大学）

1. われわれは感情をどう考えているか

感情がいったい何であるのか、という問題にアプローチするためにわれわれが取った方法は、感情の観察でも、感情理論の研究でもなく、感情の作成である。もちろん、感情という現象は高度な認知的現象であるから、われわれの目標をフルヴァージョンの感情の作成におくことはできない。そのような感情は、たぶん、フルヴァージョンの認知的行動者、つまり完成体に近いロボットしか持ちえないであろうからだ。したがって、われわれが作成の対象としたのは、極めてプリミティブな段階の感情と目されるものである。そうした感情の一つとして、われわれは＜共感＞を取り上げ、それを、ニューラルネットワーク（NN）群を訓練する過程で個々のNNが創発的に獲得する＜行動の傾向性＞として実現しようとした。当然のことながら、われわれがこのようなアプローチをしたことについては、いくつかの理論的な見込みが前提となっている。以下、まずそうした前提について述べ、今回の作成過程でわれわれが用いた2つの特別な構成手続き（レシピ）を解説し（2節）、それによってどのようにNN群が訓練され、どのような行動傾向を実際に獲得していったのかを説明する（3節）。そして最後に、NN群が獲得した行動傾向がわれわれの期待する感情機能を示すことの検証として、「囚人のジレンマ」課題を課したことの理由とその結果を、今後の見通しとともに述べる（4節）。

ところで、感情の本質に関して恐らくもっとも気になることに、ここで触れておこう。それはクオリアの問題である。この研究は、平成16年度から18年度にかけての科学研究費課題「意識と感情をもつ認知システムについての哲学的研究」（代表者・柴田正良）の一端であるが、それは最終的には、意識とクオリアに関するいわゆるハードプロブレムの存在論的・

認識論的な解決を目指すものである。しかし、今回の研究の射程からは、クオリアに関するそうした問題はさしあたり除外されている。われわれがターゲットにしたのは、「悲しく感じられるとはまさにそれであるような、そのそれとは何か」といった感情の主観的・現象的な経験の質ではなく、客観的な因果過程の中で第三者的に観察可能な感情の機能に対する問いである。現象を理解するためのもっとも効果的な手段としてその現象を作るつもりなら、機能ではなくクオリアをこそ作るべきだったろう。しかし、われわれはそうはしなかった。なぜクオリアではなく機能なのか、ということの答えは、簡単でもありまた難しくもある。ともかく、一つの明確な理由は、われわれの誰にとっても、クオリアを人工的に作り出すことなど現時点ではとうてい出来ないということだ。それはおよそ何かバカげたことのように思われる。しかし他方、この不可能性が概念的なものなのかどうか、クオリア概念は見かけにもかかわらず混乱したものでないのかどうか、そしてもしクオリア概念の適正な改変がありうるならば、その下ではクオリアの人工的な生成が可能となるのかどうか、といった一連のことには、もつれにもつれた哲学的問題がある。したがって、クオリアではなく機能に向かったわれわれの作業は、機能からクオリアへの帰路がまだ概念的に明確でないという意味で、一刀両断には語ることでできない不安定な部分を含んでいる。一つの別の言い方をすれば、クオリアの人工的な生成にはしかるべき機能の実現で十分だ、という概念的可能性ですらまだ残されているのである。

いずれにせよ、われわれは一気に意識とクオリアのハードプロブレムを攻略しようとは思わなかった。われわれが人工的に作成しようとしたのは、クオリア抜ききの感情の機能である。したがって、避けがたいことではあるが、われわれがどこまで感情機能を正確に模倣しようとも、「それは本当に感情なのか？」という懐疑に常につきまといわれることになる。正直に言えば、研究チームの中においてさえ、この懐疑は今もくすぶっている。これに簡単に応える術はない。しかし、少なくともわれわれ著者のうちの一人は、この懐疑が、初めて鏡の中に自分の姿を認めたチンパンジーの、「これは本当にわたしなのか」という疑い（必ず彼が抱くであろう当惑）と同じようなものだと考えている。要するに、原理的な問題であるよりは、思考や生活スタイルの「慣れ」の問題なのだ。だが、このように機能とクオリアの関係を理解するためにも、まだ多くの議論が必要であり、その成否を含めてわれわれの今後の課題としなければならない。

われわれは感情を、行為者の基本的認知機能を飾るフリルだとは考えない。感情は、知覚、記憶、判断、意思形成などの認知機能と並んで、それ独自の機能を果たす重要な認知モジュールである。むしろ、感情こそ、他の認知機能に従わせる認知の主役であるとも考えられるが、しかし、感情機能はおそらく進化のプロセスにおいて、次第に複雑化する認知・行動機構をいわばくそのつど事後的に>制御するものとして形成されてきたという意味で、一枚岩の機能ではなく、多層のもしくはモザイク的な機能集合体であろう。そのことは、感情が、身体 of 生化学的な状態と直結する内臓・皮膚感覚ベースの次元から、高度に慣習化された社会でのみ発現する社交的次元までを含む、広範囲に渡る現象であることから理解される。

今回のわれわれの感情NN群は、この広範囲、多次元に渡る感情機能のうち、プリミティブな層にある一つの感情のシミュレーションをめざしている。その感情は、共感、つまり他者との協調的な行動を率先して取るためにある程度 of リスクを自らに引き受ける、という行動を発現する内部的な状態である。「プリミティブな感情」というわれわれの表現が誤解を招かないように言うておけば、われわれは、「情動 emotions」という言葉を感情の最深部の基層にあり、身体ベース of 生化学的次元と融合する感情を意味するために用いる。それに対し狭義の「感情 feelings」は、「情動」に反し、あるいは「情動」に荷担しながら、「情動」とは別の機能を果たす（われわれは「情念」を、広義の「感情」とほぼ同義と解する）。そこで、われわれの言うプリミティブな感情とは、狭義の感情であるがゆえに、後に見るように「恐怖」という情動に逆らって独自の機能を果たすが、それ自体は、情動に最も近い感情の古層に存在する部分だ、ということになる。

それでは、われわれの感情作成における理論的な2つの前提を手短に述べておこう。その一つは、感情が機能であるとしても、その機能は計算論的に<厄介な intractable>ものだということである。もう一つは、われわれのターゲットにした共感感情はとくに、社会的感情として集団内で進化してきた、ということである。

第一の点は、感情機能のシミュレーションを行うに際して、われわれに古典的なプログラムではなくニューラルネットワーク（NN）を採用することを選択させた。ここで、昔から感情が理性によつては制御しがたい、非合理的な衝動としてよく描かれていることを思い出そう。われわれは、このことに一片 of 真実以上のものが含まれていると考える。つまり、それ

は、感情が機能をまったく持たない生物学的な擾乱ではないとしても、計算という意味での理性的な機能を果たしているわけではない、ということだ。しかし、逆にこのことは、われわれにとって感情が厄介な存在であるどころか、極めて生存価値の高い機能を＜計算によらずに＞果たしている、という可能性を示唆するものである。

人工知能の短い歴史の中でも「フレーム問題」という難問は、計算的合理性の限界を鮮明に示したものとして名高い。「計算量の爆発か、記憶量の爆発か」というフレーズで知られているこの問題は、人工知能にとっては、いわゆる計算論的な問題解決法のジレンマとして実現されるが、われわれ「自然知性」にとっては、何らかの仕方ですそのジレンマが解決もしくは大幅に緩和されているように見える。われわれがチューリング・マシンという抽象的な数学的存在ではなく、生物学的な次元で具体的に受肉されていることを考えるなら、何人かの論者ととも、フレーム問題を解決する＜計算なき合理性＞が生物学的なプロセスと関わりのある何らかの因果的合理性によって実現されている、と想定するのはそれなりに理に適っているだろう(Cf. Oatley & Johnson-Laird 1987)。

そのような生物学的な解決法が感情の機能だということは、いまだ完全には論証されていないし、本研究チームの中にも、感情を何か急場しのぎの救世主のように扱うことに対して警告を発する者もいる。感情は概念分析に耐えられるほど「筋が通っている」、というわけだ。しかし、ここでのわれわれの作業仮説としては、感情機能は計算論的に厄介であるがゆえにプログラム化不可能である。したがって、われわれが感情に期待する機能は、実地における学習の結果として出現する＜創発的な機能＞だ。つまり、もし感情機能は何らかのプログラムによって与えられたものであるなら、感情がなす働きは、新しい事例に直面しても、われわれが最初にそのプログラムに投入したものを越えることは決してないだろう。しかし、もしそれが＜創発的な機能＞なら、感情は、予想外の事態に直面したときに、われわれが期待するような柔軟で堅固な働きをなすかもしれない。それゆえ、われわれは、感情機能のシミュレーションにおいて、古典的なプログラムではなく、ニューラルネットワーク(NN)による実現を選んだのである。

われわれの感情作成のもう一つのポイントは、進化である。これでの議論からすれば、感情は、持たないよりも持つ方がよい。これはもちろん、感情が自然淘汰の中で選択されてきた生存価値の高い機能だ、ということだ。し

かし、その機能の内実は、まだ生理学的にも、心理学的にも、認知科学的にもはっきりとは解明されていない。つまり、なぜ感情は持たないより持った方がいいのか、ということの答えは、認知的行為者内部のメカニズムの点ではまだ明確にされていないということだ。しかし、他方、社会的感情と言われる部類の感情の中には、集団内における個体相互の結びつきを強固にするものがある、という点は比較的是っきりしている。つまり、感情の多くの機能のうちわれわれがターゲットにしたのは、集団生活を営む人間たちの協調行動を下支えする、内発的な傾向性としての＜共感＞である。

われわれが社会的生物である以上、集団のメンバー相互とのさまざまな社会的やりとりをうまくこなさない限り、結局は、個体も集団の死滅とともに死滅する、ということは明白である。つまり、協調行動を含めた他者とのやりとりには、食物・安全な居住・生殖パートナーなどの確保と同様に、強力な進化圧がかかっているものと考えられる。他者との関係に関しては、進化心理学などの研究から、われわれが他者の認知状態を理解し、予測するための「心の理論」を生得的に備えているらしいことが明らかになってきた。もちろん、「生得的に」ということは「進化によって」ということである。しかし、「心の理論」が信念、知識、欲求、行為、意図といった認知状態に関する知的判断のレベルで働くのに対し、＜共感＞は感情の一種である以上、それよりもさらに下のレベルで働いているとわれわれは考える。進化の文脈でいえば、＜共感＞は「心の理論」よりも古い時代に獲得された機能なのだ。

進化を本研究の主役の一つとすることに関して、われわれは、比較的簡単なストーリーを描いた。まず、NNによってシミュレーションされる感情は、＜一人の認知的行為者：対：環境＞という場面でその機能を発揮するのではなく、ある作業を行う多数の行為者間でその機能を発揮する以上、感情を実現するNNは、一つの個体としてではなく、多数の集団メンバーとして形成されなければならない。また、＜共感＞という感情がユニークなある機能を果たすということは、＜共感＞がなければ協調行動の出現が阻害されるような逆向きの何らかの力が常に働いている、ということの意味するだろう（それらの逆向きの力もまた、場合によっては、それなりの有益な機能を果たしているではあろうが）。そのようなわけで、それらの多数のNN群は、このストーリーにおいて設定されたりスクや恐怖の情動に抗して、ある作業課題を互いに協調して達成するように訓練される。

さて、感情機能が自然淘汰の対象であるということは、集団全体の命運がこれらの行為者の協調性にかかっている、ということの意味する。つまり、集団内のメンバー相互の協調性が一定のレベル以上の成果を生み出さなければ、集団とともにそれらのNNはすべて死滅するのだ。他方、一定のレベル以上の成果を上げた行為者たちのうち、優れた素質を有する者たちはそうでない者より遺伝的に優遇されるはずである。つまり、最初の作業課題において優秀な成績を収めたNNたちは、選択的に次世代にその素質を残すが、そうでないNNは自分たちの子孫を残すことができない。もちろん進化のシミュレーションにおいても、われわれは、獲得形質（学習の結果）が遺伝することになるラマルク型の進化ではなく、優れた学習結果を残した素質が遺伝するようなダーウィン型の進化を選択した。この意味で、＜共感＞は、経験的学習による創発とダーウィン型進化の両方の条件の下で人類が獲得してきた感情だ、というのが本研究におけるわれわれの二つの前提の内容である。

2. われわれのレシピ：ニューラルネットワークと遺伝アルゴリズム

2.1 構成

以上のような意味で、人類は進化の過程を通して共感性の感情を獲得した。このことのコンピュータ・シミュレーションを具体的に述べよう。ある村（集落）の人間、つまりハンターたちが、マンモス等の獲物を獲得するという協調作業を幾世代にも渡って行う。その村は、何世代か後には、マンモス等の獲物の獲得に成功して存続しているか、獲物の獲得に失敗して死滅しているか、のいずれかの状態になる。獲物の獲得に成功して存続しているならば、その村のハンターたちは共感性の感情を獲得しているであろうし、共感性の感情を獲得していないならば、その村のハンターたちは獲物の獲得に失敗して死滅しているだろう。

上記の事柄を人工ニューラルネットワークと遺伝アルゴリズムでコンピュータ・シミュレーションした。コンピュータ・シミュレーションの詳細は以下の通り。

ハンター→ニューラルネットワーク

村（＝ハンターの集合）→ニューラルネットワークの集合

各ハンターの獲物獲得の学習→ニューラルネットワークの誤差逆伝播
(Back Propagation: BP) 学習
ある世代からの次世代の生成→ 遺伝アルゴリズム(Genetic Algorithms:
GA)

そこで、まず、ニューラルネットワークと遺伝アルゴリズムについて以下
で簡単に説明する(Tsukimoto & Hatano 2003), (森田 2005)。

2.2 ニューラルネットワーク

ニューラルネットワーク (NN) には、種々のタイプがある。本研究で
使うのはフィードフォワード型ニューラルネットワーク(Feed-Forward
Neural Network)である。各層のニューロンの出力がそれぞれ次の層のす
べてのニューロンに繋がっており、信号の流れは一方向である。3層が最
もよく使われているが、その理由は、情報処理では3層で十分だからであ
る。

3層のニューラルネットワークの例は図1の通りである。3層は、左から
入力層、中間層、出力層と呼ばれる。中間層は隠れ層ともいう。図1の場
合は、入力層の素子は2個で、中間層の素子は4個、出力層の素子は1個で
ある。

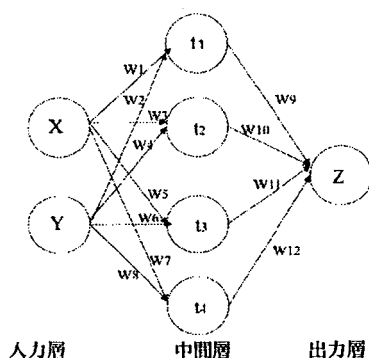


図1 ニューラルネットワーク

各々の素子を式で書くと、

$$t_1 = S(w_1x + w_2y - h_1)$$

$$t_2 = S(w_3x + w_4y - h_2)$$

$$t_3 = S(w_5x + w_6y - h_3)$$

$$t_4 = S(w_7x + w_8y - h_4)$$

$$z = S(w_9t_1 + w_{10}t_2 + w_{11}t_3 + w_{12}t_4 - h_5)$$

である。 w_i は重み係数と呼ばれ、 h_i はバイアスと呼ばれる。 S は出力関数と呼ばれる。いくつかの出力関数があるが、最も良く使われるのは、シグモイド関数 $S(x) = 1/(1 + \exp(-x))$ である。シグモイド関数は図2のような非線形関数である。

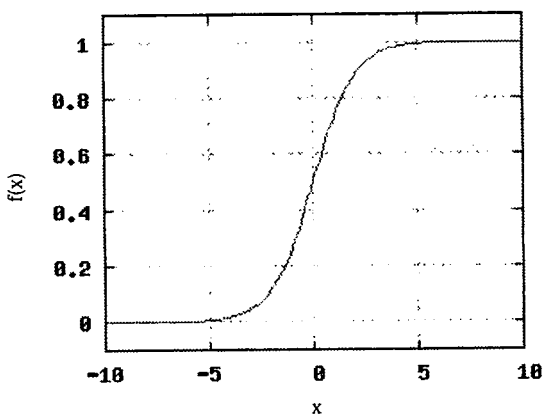


図2 シグモイド関数

上記のようなニューラルネットワークの学習の代表的な方法に誤差逆伝播法という方法があるが、それは、最初に適当に重み係数を決め、誤差が小さくなるように重み係数を反復して修正するという方法である。以下に概略手順を示す。

1. 最初に、適当に(通常は乱数を使って確率的に)、重み係数とバイアスを決める。
2. その重み係数とバイアスを用いて、入力(図1では、 x と y)から出力(z)を計算する。(順方向)
3. 2.で計算された出力(z)と教師信号の差(誤差)を求める。

4. その誤差を小さくするように、出力層から入力層に向かって、各素子の重み係数とバイアスを修正する。(逆方向)
5. 上記2, 3, 4を反復する。
6. 誤差が一定値以下になるか、反復回数が一定値以上になれば、停止する。

2.3 遺伝アルゴリズム

遺伝アルゴリズム(Genetic Algorithms:GA)は、遺伝による進化をコンピュータ・シミュレーションするものである。GAでは、複数の個体で構成された集団に対し、ある問題に対して優れた解を導き出す形質を持った個体を残すように、個体の遺伝子に対して遺伝子操作を何世代にも渡って繰り返し適用する。集団の個体数は一定の場合が多い。

基本的には、優れた個体のみに次世代の子孫を作らせる。より具体的には、なんらかの適応度関数を設け、その関数の値が大きい個体(=優れた個体)を二つ選択し、その個体の遺伝子に対して、交叉と突然変異の操作を行い、次世代の個体を(複数)生成する。遺伝子は通常、0/1のビットが用いられるが、実数や記号で表現される場合もある。工学的には、なんらかの評価関数に対する優れた解(=個体)を確率的に探索する一手法である。

GAのフローチャートを図3に示す。

・遺伝子(表現)について

本研究では、ニューラルネットワーク(個体)の重み係数(とバイアス)の2値表現を遺伝子と考える(図5参照)。

・選択について

次世代を作れる個体を選択する方法にはいくつかの方法があるが、本研究ではルーレット法を採用する。ルーレット法の基本的な考えは、各個体は、その適応度関数の値の大きさに比例して、次世代生成に関与できるというものである。

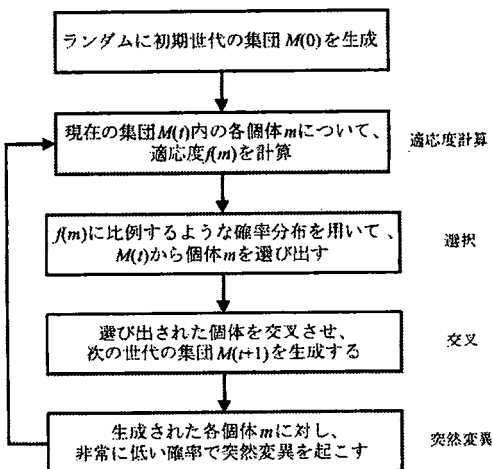


図3 GAのフローチャート

例えば、個体数が5で、その適応度関数の値が、 $f_1, \dots, f_5$ であったとしよう（図4参照）。この場合には、4番目の個体が一番高い確率で関与でき、3番目の個体が一番小さい確率で関与することになる。

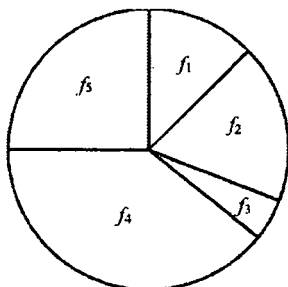


図4 ルーレット法

・遺伝子操作について

遺伝子操作には、交叉と突然変異がある。最も基本的な交叉は以下の一点交叉である（図5を参照）。図5の例では、左側の二つの2値表現された数字（01列）に対して、ある点以降の01列を入れ替える。入れ替えた後が右側である。

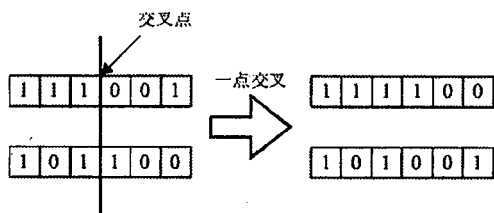


図5 交叉

突然変異に関しては、ある確率で、あるビットに対して、0から1（もしくは1から0）の操作を行う。

2.4 GA+NNについて

GAとNNの組み合わせ方法としては、NNの重み係数に対してGAの操作を行うことで、重み係数を最適化するという方法が多い。この方法は、工学的な最適化問題には適切かもしれないが、進化による感情獲得のコン

ピュータ・シミュレーションをめざす認知科学的（もしくは心理学的）な本研究の場合には、不適切である。なぜならば、優秀な個体を適応度で選択するが、その適応度(関数)を学習(最適化)する個体学習もGAで行うことになるからである。そこで、本研究では、遺伝による進化(集団学習)と、個体であるNNの学習(個体学習)とを分けて考え、遺伝による進化はGAで行い、個体学習は誤差逆伝播法で行う。

誤差逆伝播法でたどり着く極小点（誤差の極小点）は、それぞれの極小点ごとに適応度を有しているが、これは、その極小点を有する引き込み領域の適応度とも言える。したがって、本研究の方法は、適応度が高い極小点（引き込み領域）を遺伝的アルゴリズムで確率的に探索する方法と言える。誤差逆伝播法に関しては、引き込み領域の初期状態から極小点までの移動を行う。

上述の方法においては、ニューラルネットワークの集団における遺伝的アルゴリズムによって各ニューラルネットワークの重み係数の初期値が決まり、個々のニューラルネットワークは、与えられた初期値から誤差逆伝播法で与えられたデータを学習する。そして、選択（淘汰）によって、適応度が高いニューラルネットワークだけが残る、それらの重み係数の交叉等で次世代のNNの重み係数の初期値が決まる。

3. ニューラルネットワーク群（ハンターたち）の実際

われわれ人類の祖先のある段階では、感情反応らしきものが見受けられても、本来の意味での協調性（共感性の感情）は出現していない、ということが想定可能であろう。そこで、やや乱暴ではあるが、われわれの背景仮説として、彼らには情動はあるが感情はない、と仮定しておくことにする。

さて、われわれの獲物獲得ゲームの狙いは、次のような感情進化をシミュレーションすることである。すなわち、最初の世代の「人類」は、情動のみを備えていてまだ共感性の感情は備えていないが、その幾世代も後の「人類」、すなわち獲物獲得に成功して存続している村のハンターたちは、情動に加えて共感をも獲得している。したがって、協調行動が獲物の獲得を可能にするという意味で共感が生存の必要条件となるためには、獲物を一人のハンターでは倒せず、複数のハンターでのみ倒せるように設定しておく必要がある。

獲物獲得ゲームの画面を図6に示す(森田2005)。

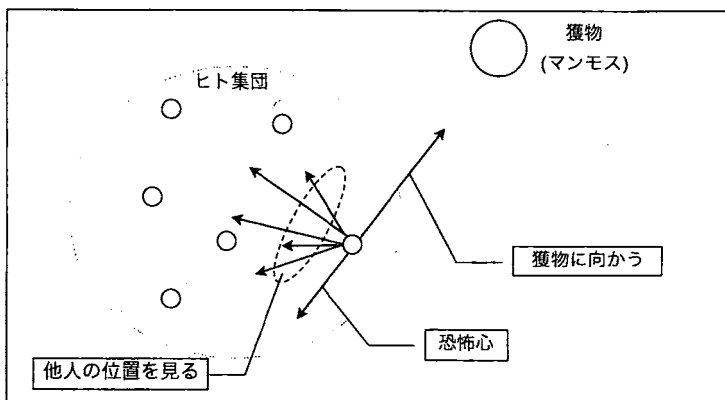


図6 獲物獲得ゲーム

われわれの設定条件では、村人であるハンターたち数十人が、お互いの位置を見ながら、獲物に近づいてゆく。獲物から一定の距離以内に、同時に複数人（例えば3人以上）が入れば、その獲物を倒せる。同時に1人か2人しか入れなければ、その獲物を獲得できず、負傷する。負傷すると、次の獲物獲得には（基本的には）参加できない。

各ハンターは、ニューラルネットワークでシミュレーションされる。各ハンターが獲物に近づくメカニズムは、獲物との距離をBP学習で近づけることで実現する。誤差逆伝播学習（BP学習）での一反復は、獲物獲得で獲物に近づくステップ（歩）に対応させる。したがって、あるハンターが、20歩で獲物にたどり着いたならば、それは、BP学習での20回の反復で目標関数の値を0にしたことを意味する。

各ニューラルネットワークの入力は、全ハンターの(xy)座標である。例えば、20人のハンターがいれば、入力数は40になる。各ニューラルネットワークの出力は自分の次の位置を示す(xy)座標であり、出力数は二つである。

BP学習の目標関数は、獲物との距離だけではなく、それに、恐怖（情動）と、協調性（共感性の感情）を考慮したものにしておかなければならない。もし、目標関数に、恐怖と共感が無ければ、BP学習は、単に獲物に近づけばよいので、きわめて簡単なものになり、数回の反復で終わる。具体的には、NNは次の式が最小となるようにBP学習を行う。

$$\left\{ \begin{array}{l} \sqrt{(x_i - t_x)^2} + \lambda_1 \sum_{j=1}^n \sqrt{(x_i - x_{(j)})} + \lambda_2 \sqrt{(x_i - t_{(x)})} \\ \sqrt{(y_i - t_y)^2} + \lambda_1 \sum_{j=1}^n \sqrt{(y_i - y_{(j)})} + \lambda_2 \sqrt{(y_i - t_{(y)})} \end{array} \right.$$

上式において、 x_i, y_i は*i*番目のハンターの座標、 $x_{(i)}, y_{(i)}$ は*i*番目以外のハンターの座標、 t_x, t_y は獲物の座標、 n はハンター集団の数、 $t_{(x)}, t_{(y)}$ は獲物と逆の方向(逃げる方向)の座標を表している。第1項は獲物とハンターとの距離、第2項は他のハンターとの距離の総和、第3項は逃げる方向の距離を示すものであり、 λ_1 は共感性を、 λ_2 は恐怖心を表現する0.0以上1.0以下の実数である。

ここで、一つ注意してもらいたいのは、恐怖(情動)も、共感(感情)も、このシミュレーションでは、BP学習の目標関数の中に組み込まれていて、外在的な場として表現されている、ということである。したがって、各ハンター(ニューラルネットワーク)の個体としての特徴とは一連の重み係数に他ならないが、各ハンターの情動や感情は、目標関数として場的に表現された逃走行動(恐怖)や協調行動(共感)に関するBP学習を通して、各ハンターの内面であるところのその重み係数として獲得されるのである。

第一世代の各ハンターの最初の重み係数、つまり獲物獲得開始前のニューラルネットワークの重み係数は、乱数で決める。この重み係数には、恐怖(情動)も共感(感情)も反映されていない。われわれの仮定では、恐怖(情動)は、この段階ですでに進化によって獲得されたもの(感情より古層の機能)であるが、その恐怖(情動)を、内在的にニューラルネットワークの重み係数で表現するのは困難なので、外在的に表現することとした。

第一世代のハンター(ニューラルネットワーク)の重み係数は、獲物獲得を通して変化してゆく。数十回獲物獲得を行った時点で、次世代のハンターを生成して世代を変える。これを何世代にも渡って行う。次世代を生成するときには、適応度の高い個体を選んで遺伝子操作を行うが、適応度は基本的に獲物獲得に参加した回数である。

進化の一例は図7の通りである。横軸は世代数であり、縦軸は一世代あたりの獲得した獲物数である。徐々に、獲物数が増えていることが見て取れる。

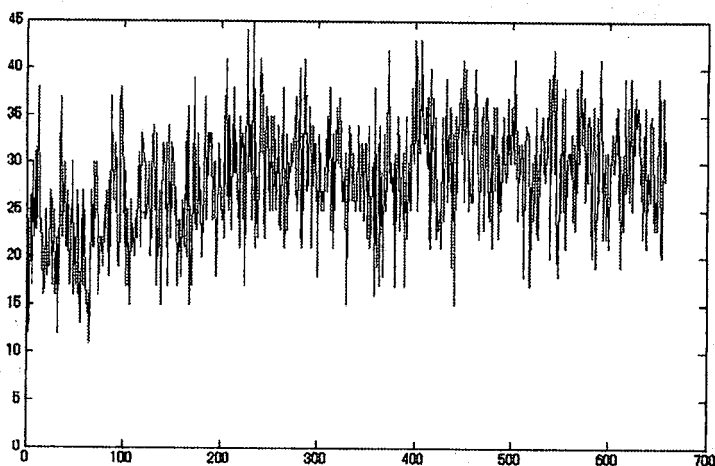


図7 進化の一例

4. 検証課題と今後の見通しについて

獲物獲得ゲームでNNが感情を獲得したか否かの確認として、進化させた世代から適応度に基づき個体を取り出し、別の課題に挑戦させる、というものが考えられる。本研究では、その課題として囚人のジレンマ・ゲーム（獲物獲得版）を選んだ(森田2005)。

囚人のジレンマ・ゲームとは、2人の囚人A、Bが共に利他的な協調行動（黙秘）を行った時の2人の利得の合計が最大(受ける懲役年数が最小)となるように設定された繰り返しゲームである。囚人のジレンマ・ゲームにおける利得の例を表1に示す。表の数字は受ける懲役年数を示しているため、数値が低いほど高い利得となる。

表1 囚人のジレンマ・ゲームの利得の例

	B: 黙秘	B: 自白
A: 黙秘	A(2) / B(2)	A(15) / B(1)
A: 自白	A(1) / B(15)	A(10) / B(10)

NNに囚人のジレンマ・ゲーム（獲物獲得版）を行わせる場合、利得は表2のようにする。囚人のジレンマ・ゲームと違い、数値が大きいほど高

い利得となる。集団数は2で、捕獲は2人の場合のみ可能とする。例えば、Aが「捕獲」(有効半径内に侵入)しても、Bが「捕獲せず」(有効半径外)ならば、Aは負傷者となり、少ない利得(1)しか得られない。

表2 囚人のジレンマゲーム(獲物獲得版)の利得の例

	B:捕獲	B:捕獲せず
A:捕獲	A(10) / B(10)	A(1) / B(15)
A:捕獲せず	A(15) / B(1)	A(3) / B(3)

このような囚人のジレンマ・ゲームの改訂版を本モデルで進化させたNNに行わせることによって、NNが感情を獲得したか否かを検証した。ただし、上記の表2のゲームを行うNNには、自分の座標しか入力しない、すなわち、もう一人の座標を入力しない。これが、獲物獲得ゲームとの重要な相違点である。獲物獲得ゲームでは、他人の座標も入力して誤差逆伝播学習を行った。しかし検証課題では、自分の座標だけで他人の座標を入力せずに、A、Bの2名がどれだけ協調性があるのかを見るのである。

上記の検証課題を行った結果、適応度が高い二つの個体の方が、適応度の低い二つの個体に比べて、利得の合計が高いことが確認された。例えば、表2の課題を100回行った場合、適応度52の2個体では、558点と460点を得たのに対し、適応度0.9の2個体では、336点と336点しか得られなかった。

さて、検証課題のこの結果をどう評価すべきだろうか。われわれのNNが獲得したものは感情機能だ、言えるのだろうか。あるいは言えるとして、どのような意味においてなのだろうか。われわれのNNが協調行動を誘発するある種の内部状態を獲得したのは間違いない。したがって、その内部状態が感情であると言うためには、少なくとも2つのレベルで検証が行われなければならないだろう。第一は、この<共感>がさまざまな作業課題で同様の協調行動をNN同士の間で生み出す、ということの検証。そして第二は、これが、NNの個体内部で<共感>に期待されるようなある種の認知的機能を果たす、ということの検証である。

われわれが囚人のジレンマ・ゲームを検証課題に選んだ理由は、まさに第一の意味においてである。もちろん、最初の獲物獲得ゲームと検証課題が基本的に「同じもの」であったなら、NNの共感能力は検証されたことにはならないだろう。少なくとも、同じ状況で同じ行動をなす、というだ

けのことはどのNNも学習済みだからだ。だが他方で、検証課題の課す状況がNNの共感能力にとって基本的に「無縁なもの」であったなら、それもまたNNの能力を検証することにはならないだろう。この意味で、最初の獲物獲得ゲームと囚人のジレンマ・ゲームとの異同はやや微妙に感じられるかもしれない。

ここで、囚人のジレンマについての詳細な解釈を述べることは控えるが、われわれの観点からすれば、囚人のジレンマは、短期的な利己的戦略によっては合理的な解決が望めないパズルの実例である。つまり、もし短期的に自己の利得だけを考えるなら、囚人はいずれも「自白」することを選び、その結果、両者は<集団としてみれば>最悪の利得（最大の懲役年数：合計20年）を得ることになるであろう。したがって、個人としての短期的な合理的戦略は、必ずしも<集団として>最善の結果をもたらさないのだ。それゆえ、われわれの文脈に即して言えば、囚人たちが「母なる自然」から最大の利得を引き出すためには、一定のリスクにもかかわらず、長期的な視点に立った利他的な戦略である「黙秘」を採用しなければならない。

しかし、このような長期的な視点に立った利他的な戦略は、進化の途上においては、人類の合理的な思考の産物であるよりは、他者に対する生得的な態度の結果である方が確からしいであろう。われわれの祖先は、昔からそれほどには経済学者でも哲学者でもないからだ。したがって、われわれは、他者に対する<共感性の感情>を獲得したNNたちこそが、この検証課題を、他のNNの行動に関する情報なしに、自己の生得的な傾向性として高いレベルでクリアすることを期待したのである。先に述べた、獲物獲得ゲームとの「重要な相違」とはこのような意味であり、この検証結果に関しては、われわれは満足している。というのも、われわれのNNこそ、そのような利他的な素質と能力を学習と自然淘汰によって獲得したはずだからである。

しかし、囚人のジレンマ・ゲームは、なお、われわれにとっての最も望ましい検証だというわけではない。検証の第二のレベルを思い出そう。もしわれわれの見込み通りに、感情がかなり高度な認知的機能を果たしているなら、それは、認知的行為者としての一つのNNの認知活動の中で発揮されていなければならない。したがって、NNの獲得した<共感>能力が本物なら、それは、獲物獲得ゲームとはかなり異なる状況、例えば対象知覚や言語解析といった認知的課題においてもその機能を働かせるはずであ

る。もちろん、われわれのニューロン活動の本質、つまりNNの表象能力が万能シミュレーション能力にあるとしても、われわれの脳の機能局在が如実に示しているように、巨大な一つのNNがわれわれの認知のすべてをこなしているわけではない。したがって、NNによってシミュレーションされた感情機能が、あらゆる認知的難問を解決するわけではない。しかし、それにしても、われわれのNNの獲得した機能が感情機能であるなら、われわれが感情の機能に見込んでいるだけの高次の認知的機能をワイルドな状況の中で果たすべきである。

さて、こうして、われわれは、次の課題に導かれる。人工のNNが現実のワイルドな認知的状況の中で機能を発揮することは、現在のコンピュータ・シミュレーションでは原理的な限界にぶつかる。そもそも、さまざまな知覚課題やコミュニケーション課題の表現を、われわれのNNにどのように入力できるというのだろうか。われわれは、これまでネットワークは空間認知と言語解析のような本当に別々の課題を同時にこなすことはない、というカミロフ・スミス的不满を自らに引き受けなければならない、という状況にある(Karmiloff-Smith 1992, pp.180-81)。つまり、認知のシミュレーションはあくまでシミュレーションに過ぎず、認知そのものではないのだ。それはちょうど、たった一つの信念しか持たない認知的行為者は、実はそもそも信念をまったく持っていないのと同様である。したがって、われわれが次になすべき検証は、われわれのNNを、ロボットのような実在の世界におけるトータルな認知的行為者に搭載し、他の認知能力との協同関係の中で、その感情機能を検証することである。だがしかし、すでにお気づきのように、この検証は、もはやわれわれにとっての当初の検証課題というより、それを越えて、新たにトータルな認知的行為者を作り上げることに他ならない。それは、われわれが今後、クオリアと意識についての概念的な解明と往還的になすべき、認知ロボティクスの課題である。

謝辞： NNの学習プログラムの作成およびシミュレーション作業に関して、東京電機大学の大学院生、森田晋一郎氏と恩田宣和氏に感謝いたします。

参考文献

Karmiloff-Smith, A., *Beyond Modularity*, The MIT Press, 1992, (『人間発達の認知科学：精神のモジュール性を越えて』, 小島康次・小林好和訳, ミネルヴァ書房)

キーナン, J. P. 他, 『うめばれる脳—「鏡のなかの顔」と自己意識』, 日本放送出版協会, 2006.

森田晋一郎, 「生物学的手法によるニューラルネットワークの進化・学習」, 東京電機大学修士論文, 2005.

Oatley, K., & Johnson-Laird, P. N., "Towards a Cognitive Theory of Emotions", *Cognition and Emotion*, 1, 1987.

柴田正良, 「ロボットがフレーム問題に悩まなくなる日」, 『シリーズ心の哲学Ⅱ: ロボット篇』 (勁草書房: 信原幸弘編) 所収, pp.119-174, 2004.

Trappl, R., & Petta, P., et al.(eds.), *Emotions in Humans and Artifacts*, The MIT Press, 2002.

Tsukimoto, H. & Hatano, H., The Functional Localization of Neural Networks Using Genetic Algorithms, *Neural Networks*, Vol.16, pp.55-67, 2003.