人工知能 課題番号 25「課題図書を読んで人工知能、人工生命の実現可能性について論じよ。」

工学部電子情報工学科 03-175001 浅井明里

2018 年 1 月 3 日

1　本課題について

本課題では、有田隆也著『心はプログラミングできるか』の要約及び主旨についてまとめた上で、この書籍全体で取り上げられている「人工生命」と「人工知能」の実現可能性、及び「心とはプログラミング可能か」という問いについて論じていく。

2　「心はプログラミングできるか」の要約及び主題

著者はまず前半の 4 章で、人工生命や創発現象に関わる代表的ないくつかのトピックを紹介している。

第 1 章では「蟻の餌集め」などを例に挙げ、「一匹一匹の頭はよくない」蟻が、群になるとそれぞれの蟻から想像できないような知的な処理を蟻の間の相互作用で実現できるという「群知能」について論じている。まず紹介されている「蟻のフェロモン現象」とは蟻は餌を見つけた時フェロモンを発することにより、他の個体に餌の在処を伝え、また同時にこのフェロモンは時間と共に拡散・蒸発するために、現在もしくは現在に近い過去に餌があるという情報も伝えるというものであり、言い換えれば、このフェロモンは「場所」と「時間」という二つの概念を同時に伝えている。同じ餌に至る長い道と短い道が存在する時、短い道の方が往復する個体の数が増加するために、フェロモンの濃度が濃くなり、故により多くの個体を誘引しやすくなるという「正のフィードバック」がかかり、結果として集団としての蟻がより効率よく餌を獲得することを手助けしている。実際にこの蟻の餌集めの原理を利用したサウスウェスト航空は、問題であった貨物運送の効率の悪さを「蟻の餌集め」からヒントをあつめ、ルーティングの方法を改善した、結果として、フライト数は 80% も現象し、人の作業量は 20% も減った。この結果削減されたコストは 1000 万ドルにも登るという。こういった餌集めだけでなく、分業の仕方や大きな餌の運搬なども蟻の群知能の例として挙げられている。

第 2 章では、シンプルだが生命性を感じるグラフィクスを進化のメカニズムを借りて作成するモデルであるバイオモルフを例に挙げ、「進化のメカニズムと人の感性の共同作業」について論じている。バイオモルフでは人がどの個体を残すのかを主観で選ぶという点が遺伝的アルゴリズムなどと異なり、人間の主観的な選択による小さな突然変異が積み重なることにより、人が進化の方向性をこまめにコントロールしつつも、最終的には意外な結果が得られることがビジュアルに実感できる。こういった遺伝的アルゴリズムにおける適応度評価の部分を人間の主観的な好き嫌いで行うようにしたものを「対話型進化」という。筆者はこの対話型進化を興味深いものとしつつも、これを進化の産物だけ楽しむのではなく、オープンエンドな進化として楽しべきではないのだろうかと提唱している。

第 3 章では、デジタル生命の元祖となった「ティエラ」と現在最も活躍している「アヴィーダ」を取り上げ、デジタル生命研究の基本から最先端を論じている。1990 年にジョン・キャスティにより作成されたティエラは仮想マシン上で動く機械語のプログラムとして「デジタル生命」であり、基本的にはただ自分と同じプログラムをコピーし続け、タイムシェアリング方式でこれらのプログラムを実行するものであった。しかし突然変異（機械語命令を勝手に変更してしまうこと）を仮定したことで、この突然変異が計算機の中でティエラに「進化」を起こさせ、のちの人工生命研究の礎となった。ティエラは自己複製を効率的に行うことのできるプログラムが進化によって探索され、複雑な相互作用と創発し、また複雑な相互作用を行わずに効率的に自己複製を可能とするループアンローチングと呼ばれる技法も創発した。アヴィーダは最新の人工生命であり、「明示的な報酬を付加的に設定できる」「プログラム同士の相互作用をやや限定し、適応進化が実現しやすくなった」という点でティエラと異なる。アヴィーダを使った進化研究は複雑な形質を生み出す時は時に「山を降り

る」（適応度を下げる）ことも必要であるなど、その後様々な研究成果を世に残した。

第 4 章は創発主義に基づく人工知能の主張について、これまでの議論が展開された上で、人工生命の基づく構成的手法は、従来のオーソドックスな科学（仮説をたて、実験などの手段によって現実と会うかを検証する、合っていなければ仮説が間違いなので仮説を修正する）とは色合いが違うが、人工生命のアプローチは「生命とは何か」「進化とは何か」などの本質的な問いかけに対してこそ、真価を発揮すると考えているという筆者の主張が展開されている。

前半 4 章が「人工生命」に対して様々な事例や先行研究の概要が示されていたのに対し、後半 4 章では人の心も持つ様々な特性に対するこれまでのアプローチが展開されている。

第 5 章では、他の個体を助けて自分が損をするような行為は遺伝子の生き残りの観点からは理解しがたいにも関わらず、なぜ自分にとってマイナスとなるような犠牲を払って他の個体にとってプラスとなるような行動をする「利他的行為」が生命の様々なフェーズで観測できるのかについて、これまでの代表的な説と共に考察がなされている。第一の仮説として挙げられている「血縁選択説」は、近親者に親切にすることは自分と同じ遺伝子を残すことにつながるとするものであり、例えば自分が犠牲になっても自分の三人の子供が生き残るならば結果として自分の遺伝子は残されるためにあえて自分を犠牲にするという選択肢を取るであろうとする仮説であり、故に血縁の遠い個体に対しては利他的行動を取りにくいとする。第二の仮説である「互恵主義説」は見返りが期待できるならば親切をする方が徳をするため、個体は利他的行動を取るとする仮説であり、さらに利他的行為をしてあげた相手から見返りが戻ってくるとする「直接的互恵主義」とその相手だけに限らず、第三者から見返りが返ってくるとする「間接的互恵主義」に分けられる。第三の仮説は「マルチレベル選択説」であり、これは多様なグループに別れているならば親切をし合っているグループの個体の方が有利になるとする仮説である。これは付き合いをするグループに利他的行動を取る「お人好し」と利他的行動を取らない「わがまま」が混在するという多様性があり、これ自体は脆いため放っておくとすぐに「わがまま」だけになってしまうが、各個体が自分のいるグループの居心地が悪くなると属しているグループを飛び出しやすくなるというモデル (環境応答移住) により、「お人好し」な個体が完全には全滅することなく様々なグループに移動することにより存続すると仮定する。このいずれの仮説でも、協調関係とは不安定で崩壊の危機にあるが、ユニット間の自発的な流動性によりユニット間のばらつきがらもたれ、ユニットを包む上のレベルのユニットは協調的に機能を発揮して存続しうるという緊張感に満ちた協調関係を前提としている。

第 6 章では、長いスケール、つまり集団において世代交代を何度も繰り返すようなレベルで起こる適応である「進化」と、短い時間スケールである個体の生涯レベルで起こる適応である「学習」という二つのメカニズム間の関係について考察している。かつては「適応度地形」の考え方からもわかるように、各生物個体の適応度地形上の位置は遺伝的情報によって生まれつき定まったもの（進化）であり、変化しないものであると考えられてきたが、人間ほどの知性をもつ生物個体の形質を全て遺伝子で記述するのは無理があり、実際には 高等な生物ほど「表現系可塑性」が増し、生まれ育つ環境に応じて、それぞれが遺伝的情報で規定されているものから離れていくという現象（学習）が影響を及ぼしている。この進化と学習の相互作用について、ボールドウィン効果は「表現可塑性の大きい方がより大きい適応度を達成できるという学習のメリットが働き、表現可塑性が増加し、同時に適応度は増加する」と、「到達適応度が集団で変わらなくなると、学習のデメリットが働くようになり、獲得形質は徐々に先天的形質におきかわり、適応度は高く保ちつつもゆっくりと表現可塑性は減少する」という二段階のモデルを示すものだと解釈されることが多い。この進化と学習の相互作用については全体像を把握することは容易ではないが、筆者は 社会生活を営む上で相手次第で振る舞いを変えるために必要不可欠な「心」とは表現可塑性の極地であると主張し、また言語の発達などの不連続な進化の際にはこういった表現型可塑性が重要な働きをしていたと述べる。

第 7 章では感情にアプローチする様々な人工生命的手法が紹介され、「人の感情とは何か」についての論が展開される。人間の感情については、感情の成り立ちや機能に関して身体的反応を第一義的に注目する立場とするジェームズ論、現実をこうだと認知していることが自体が感情体験に影響するという立場を取る認知説、社会や文化が感情体験を構成する社会構成説等諸説があるが、筆者は感情とは感情とは「物理的環境への適応」と「社会的感供養への適応」の二つの意味づけがあると考えている。また計算機の中の仮想敵世界で感情がどう進化するかを検証した実験では、例えば暗闇ではおどおどと不安そうに振る舞うなど、ロボット自身が創発した「感情」が行動に現れるようになった事例を紹介している。

第 8 章では人間の心の様々な特性に関する進化的基盤に焦点を合わせ、進化の産物として心を理解すること、あるは自然選択によって心がどう形成されてきたかを明らかにすることを目的とする進化心理学を取り上げ、心とは進化の産物なのか、文化が生成した装置なのかを紐解こうとする。デネットによる心の進化の四段階モデルは「この段階には心は存在しない、生息する環境の中でうまくやれる個体が残っていくダーウィン型生物」「この生物は自分の中に行動の選択肢を持つ、行動の結果から学習するスキナー型生物」「頭の中に世界像とでもいうべきものをもち、実際に行動に移る前に頭の中で自分の行動をシミュレートするポパー型生物」「様々な人工物を道具として扱うことができる。言語を扱える。この結果、言語による情報交換が他社の様々な過去の経験や知識に基づいて自分の行動を助けることになるので、知能を大きく飛躍させるグレゴリー型生物」に区分され、与えられた自然環境や地球における物理法則の元で、それらをうまく利用してうまく生活できるようになった結果が心や知能という立場をとる。ただこの段階モデルには、ある個体は自分だけの世界に止まっているように感じるが、筆者は、複雑化する個体間の関係、社会的環境こそが心や知能を作り上げたという立場に立ち、「ポパー型のように世界観をもち、かつその世界観の中に自分以外の他者を持つ」というモデルを示す。このモデルにおいては「心の再帰レベル」が重要になっているが、人間はちょうど適度な社会性を保つ集団でああったために、再帰レベルが高い程適応度が高くなり、その方向へ適応進化する現象が発生したために、心の再帰レベルが他の動物より深くなっていると言われている。

3　人工生命、人工知能の実現可能性について

人工生命とは、生物のような行動を研究することを目的とし、分析的な方法に代わって合成的な研究手法を用いて生命現象を探求することであり、生命のような複雑な現象をコンピュータでモデル化できれば、生物の振る舞いが的確に予測や制御ができるだけでなく、例えば蟻の餌探索アルゴリズムが航空会社の貨物の効率的な運搬の実現に貢献するなど、我々の現実世界の問題にも応用することができるようになる。

本課題で暑かった有田氏の著作を鑑みるに、自然界の生物の行動を模倣し、現実世界の最適化問題へ適用することに関しては我々はある程度成功しているのではないかと感じた。一方で、氏も述べているように、蟻のフェロモン現象一つとってもシンプルなニューラルネットワークで容易に実装できるものではなく、人手でのルールの追加や学習の監視などを行う必要があり、本来的には「自然界の相互作用により実現される」現象が完全に人手で再現できるといった段階にはまだ至っていないように感じる。またたとえ既存の生命体の単一の現象を完全に再現できたとしても、それは筆者の主張するような「人工生命」の実現とは結論づけることができないのではないかと考える。この点について考えるために、まず「還元主義」と「創発主義」について述べたいと思う。還元主義とは「複雑な現象も、それを構成するより単純な部分の性質の足し合わせとして理解できる」とする立場である。一方創発主義とは、「生命に代表される複雑なシステムにおける複雑さは、システム全体として機能したした時に初めて、構成要素間の相互作用で予期せぬ振る舞いや構造が生じる結果、創発する」とする立場である。このような創発主義に基づく人工生命について考えた場合、まだ我々は完全にはこ

の自然界における複雑な自称間の相互作用による創発を再現仕切れていないのではないかと思う。ただ、ディジタル生命である「ティエラ」と「アヴィーダ」がそれぞれコンピュータ内の仮想的な環境で独自の進化を探索し、また進化に関する新たな発見を手助けするなど、「創発現象」については少しずつ解明が進んでいるとも考えることができ、今後こういった動きは加速していくのではないかと思う。

次に「人口知能」の実現可能性、ここではこの著作の後半 4 章に寄せて「人の感情、知性といった『心』を我々はプログラム可能か」という問題について考えいきたいと思う。ジョン・サールは「強い AI によれば、コンピュータは単なる道具ではなく、正しくプログラムされたコンピュータには精神が宿るとされる」とし、人間の前認知能力を必要としない程度の問題解決や推論を行うソフトウェアを弱い AI、人間のように自意識を持ち、教師データを必要とせずとも変化していく環境を認知、把握、適応し、情報を取捨選択して記憶し問題を解決するための推論を行うようなソフトウェアを強い AI とした。著作中であげられたいくつかの例では、やはり環境だけ用意してプログラムを実行するだけでは必ずしも状況に適応した賢い知性が生まれるわけではなく、度々人手による評価関数設定やパラメータ設定等を行って初めてある程度望ましい動作をするプログラムが生成できており、まだコンピュータが自発的に学び、考え、判断するといったものからは遠く離れているように感じる。しかし一方で、ロボットが創発的に危険性の高い状況でおどおどするような行動をとるようそのロボット独自の適切な感情を習得するような結果も観測されている（実際に不確実性の高く、かつ危険性の高い状況ではこういった慎重な振る舞いに誘導する『感情』は現実世界では生命維持に有利に働く機構であるように感じられる)。

このような事例を鑑みるに、自発的に学習するソフトウェアを作成することは不可能ではないように感じられるが、私は今後強い AI を実現する上で「独立した意思・感情」を持たせることの重要性が増していくのではないかと思う。人間が時に利他的な行動をとる個人間で協調的に活動し、また個々人がそれぞれの世代での学習を通じてそれが何世代にも渡って続くことで「進化」にも影響を与えたことは人類の高度な発展に大きく貢献した。なぜ人が利他的行動をとるのか、その理由についてはまだ完全には解明されていないものの、「結果として自分の遺伝子が継承されるよう」「将来的に自分の生命維持活動にプラスに働くはずだ」という、「自分の命もしくは遺伝子をより長く存続させたい」という遺伝子レベルで規定された強烈な意思に基づくものであり、これがなければ種として高度な進化を遂げることなく絶滅してしまうのではないかと感じる。もし我々が自分で価値基準を定め、思考し、選択するようなソフトウェアを生み出そうとするならば、我々はまず「このソフトウェアは何を最終目標として存在しようとしているのか」という意思を厳密に定義する必要があるのではないだろうか。これについても「人間により価値基準を与えているのではないか」と反論されうるのかもしれないが、環境に適応し、独自の進化を遂げた生命たちがなぜ「遺伝子の存続が種の最終目標である」という遺伝子レベルでの意思を持ち得たかは不明瞭であり、ここに関しては人手による「意思」の創発を行ってもいいのではないかと感じる。

またこれに加えて、私は「心」を持つことも最終的に強い AI にとって必要な要素なのではないかと考えている。これは「恐怖」「怒り」などの感情が生命の維持もしくは協調関係を築き、結果として全体としてより生存可能性の高める（すなわち正しい選択を行う）と考えているためである。私も筆者の意見と同様に、感情とは例えば蛇を見た時に生命への危険を覚えるために「恐怖」を感じ、接近を回避しようとすることは感情のもつ「物理的環境への適応」という側面であり、また仮に複数の個体が存在しそれらの間で協調関係を構築しようとしている際に裏切ることに罪悪感を覚え、裏切りを抑制することは感情の「社会的環境への適応」を示している。「感情を持つことが人類に戦争や紛争を繰り返させ、人類存続の危機に晒してきた」とし、「感情を抑制された未来世界」を舞台としたフィクション小説、映画等は過去に多く人気を集めてきたが、やはり私はある程度の「感情」を持つことは種の存続、発展やある個体がより「正しい」選択を行うために必要な要素の

一つとなりうるのではないかと考え、また正しく推論を行う強い AI にはこういった感情の働きも重要な要素になるのではないかと考える。

ではこういった心の働き、特に感情とはどの程度までプログラムが可能なのであろうか。本書で紹介されている実験にもあるように、ある程度の環境を設定し、うまく「どういった要素が感情を規定するのか」（ロボットの実験におけるモジュレータである集中度や活性度など）を定義できれば、感情はネットワークを通してある程度学習できるのではないかと思う。一方で、もし人間が持っている程度の複雑かつ排他的でない数かぎりない感情を学習しようとした時、この「モジュレータ」をまず明確に定義することが最初の困難になりうるのではないかと思う。また時として人間は二つ以上の入り組んだ感情を持ちうる（嫉妬と羨望、怒りと失望、歓喜と焦燥等）ことがあり、こういった複数の感情の組み合わせの学習可能性を考慮すると、より複雑なニューラルネットワーク構造を定義する必要が生じるため、こちらについても容易ではないと思う。故に「感情」を人間の手によってマシンに学習させることはまだ困難であり、この点からも自分の価値判断で決断を下し、行動をする「強い AI」であるところの人工知能はまだ実現にある程度時間が必要なのではないかと思う。